THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME X (1944-45)

GENERAL EDITOR
W. B. MANN

Reproduced and printed photo-litho offset for
WM. DAWSON & SONS LTD
4 Duke Street, Manchester Square
London, W.1
1957
with the permission of The Physical Society

MADE AND PRINTED IN GREAT BRITAIN BY
D. R. HILLMAN & SONS LTD
FROME

CONTENTS

	PAGE
Preface	iv
Cosmic rays. By W. F. G. SWANN	1
Seismology. By HAROLD JEFFREYS, F.R.S.	52
X-ray diffraction techniques in the industrial laboratory. By H. P. ROOKSBY.	83
Sound. By E. G. RICHARDSON	120
Phase separation in aqueous colloidal systems. By P. KOETS	129
Physics and textiles. By A. B. D. CASSIE.	141
Some applications of physics to the processing of textiles. By J. G. MARTINDALE	172
Electron microscopy. By L. MARTON	204
The theory of the elementary particles. By H. J. BHABHA, F.R.S.	253
The mechanical design of physical instruments. By A. F. C. POLLARD	272
The lightening discharge. By J. M. MEEK and F. R. PERRY	314
Superconductivity. By K. MENDELSSOHN	358
Spectroscopy applied to molecular physics. By J. CABANNES	378
Lessons of the war for science. By J. D. BERNAL, F.R.S.	418
Author index	437

PREFACE

THE hope expressed in the preface to Volume IX that it would be possible for the current volume of these *Reports* to be the last to be published in alternate years has now been fulfilled, and it is proposed to revert to the practice of annual publication.

In the production of this volume, which has also suffered to some extent from the difficulties of war-time restrictions, I have to acknowledge again the invaluable assistance of Mr. J. H. Awbery and Dr. W. Jevons. The author index has been prepared by Dr. F. Lowater, to whom my thanks are due for a lot of very careful and painstaking work, in the course of which a great many bibliographical errors were found and corrected. I also wish to thank Dr. A. J. E. Welch for his kind help in correcting the translation from the French manuscript of Professor Cabannes' *Report*. Valuable secretarial assistance has also been rendered by Miss M. A. Webb in the Society's office.

In the current volume the modified Harvard system of references previously adopted has again been used; the author index is an index referring to authors given in the bibliographies, and is not an index of all names referred to in individual *Reports*. In the case of numeralogical references it is fairly easy to locate any author in the text; this is not, however, possible with the Harvard system without the assistance of an author index.

A corrigendum slip for insertion between pages 62 and 63 of Volume IX is being circulated with this volume.

Imperial College,*
London S.W. 7,
July 1946.

W. B. M.

* Now at National Research Council Laboratory, Chalk River, Ontario, Canada.

AUTHOR INDEX

Abrams I. M., 239, 241.
Ackemann, M., 36.
Ackermann, O., 338.
Adam, N. V., 167.
A. E. G. Forsch.-Inst., Jahrbuch, 204.
Aiya, S. V. C., 37.
Akopian, A., 327.
Alekseyevsky, N., 360, 362, 366.
Alexander, L. T., 99.
Alfvén, H., 24–26.
Allbright, J. C., 320.
Allen, A. L., 146.
Allibone, T. E., 326, 327.
Altmann, G. O., 36.
Alvarez, L., 9.
American Society of Testing Materials, 179, 182.
Ampère, A. M., 273.
Ananthakrishnan, R., 415*.
Anderson, C. D., 16, 32, 33, 35.
Anderson, T. F., 204, 239, 242, 246–248.
Angus, T. C., 144.
Angus, W. R., 395, 413.
Appleton, E. V., 349.
Appleyard, E. T. S., 366, 370, 375.
Arago, D. F. J., 284.
v. Ardenne, M., 204, 206, 217, 220, 221, 224, 226–228, 230, 232, 234, 236, 237, 242–244, 248.
Ariyama, K., 374.
Arkadiev, V. K., 372.
van Arkel, A. E., 97.
Arley, N., 22, 34.
Artingstall, F. D., 299.
Aschermann, G., 358, 360, 365.
Astbury, W. T., 95, 117, 160.
Aston, F. W., 4.
Atherton, A. L., 338.
Atkins, B. E., 148.
Atlee, Z. J., 94.
Auger, P., 8, 34, 36.
Avery, W. H., 402.

Babbage, C., 272, 273.
Babbitt, J. D., 144, 146, 368.
Bachman, C. H., 217, 231.
Baguley, P. W., 337, 346.
Bailey, C. R., 395, 413.
Bainbridge-Bell, L. H., 350.
Baker, R. F., 223, 226, 232, 236, 243–245.
Ball, R. S., 283.
Ballard, J. W., 105.
Ballaus, O., 406.
Balls, W. L., 191–193, 195.
Bangham, D. H., 158.
Bär, R., 125.
Barkas, W. W., 153, 155.
Barker, E. F., 395.
Barker, S. G., 152, 177–179, 181, 185, 187, 193.
Barnes, R. B., 248, 385.
Barnóthy, J., 45.
Barrer, R. M., 163, 166.
Barrett, C. S., 113, 115, 248.
Barry, J. G., 19.
Bastings, L., 55.
Bawden, F. C., 135.
Baxter, S., 142, 144, 145, 150, 167–169, 190.
Baylor, M. R. B., 248.
Beadle, D. G., 335.
Beams, J. W., 320, 322.
Beck, E., 337, 340, 342, 343.
Becker, R., 367.
Beer, P. L., 135.
Beger, E., 248.
Behr, A., 124, 125.
Beischer, D., 204, 224.
Beliakov, A. P., 327.
Bell, E., 337.
Bellaschi, P. L., 339.
Benedetti, S., 32, 33, 35.
Bentley, G. P., 294.
Beranek, L. L., 123.
van Bergen, W., 172.
Berger, K., 336–338.
Bergmann, P., 135.
Bernal, J. D., 135.
Bernays, P. M., 248.
Berstein, H. J., 405.
Bethe, H., 11, 20, 34, 35, 37.
Bhabha, H. J., 11, 34, 258, 264, 266, 269, 270.
Biscoe, N., 101.
Black, C. P., 144.
Blackburn, S., 161.
Blackett, P. M. S., 31, 32, 35, 258.
Blayden, H. E., 109.
Bloch, F., 38.
de Boer, J. H., 136, 156.
Boersch, H., 220, 223, 232, 237.
Bohr, N., 38.
Boner, C. P., 122.
Bonner, J., 134.
von Borries, B., 204, 213, 215, 218, 220, 222, 226, 238, 240–245.
Bothe, W., 6, 30, 34, 36.
Bouasse, H., 284.
Bouwers, A., 95.
Bowen, I. S., 2. 12.
Boys, C. V., 319.
Bozorth, R. M., 100, 116.
Brackney, H. T., 94.
Bradacs, K., 406.
Bradley, A. J., 96, 98, 107.
Bradley, R. S., 156.
Bragg, W. H., 109.
Bragg, W. L., 106, 107, 109.
Brand, W., 336.
Bray, Lt. R. J., 298.
Brester, C. J., 387.
Bristow, J. R., 366, 370, 375.
British Standards Institution, 145.
Brode, R. B., 39.
Broussard, L., 36.
Brown, K. W., 90.
Bruce, C. E. R., 325, 340–342, 345, 349.
Brüche, E., 204, 217, 218, 226, 237.
Brucksch, W. F., 361, 366.
Bruins, E. M., 8.
Brunauer, S., 156–159.
Brunt, D., 282, 284.
Buchmann, G., 124.
Buckingham, E., 302.
Büchner, E. H., 129.
Buerger, M. J., 96.
Bull, H. B., 159, 160.
Bullard, E. C., 80.
Bullen, K. E., 59, 63, 70, 73, 74, 81.
Bunn, C. W., 103.
Burghoff, H. L., 115.
Burkardt, L. A., 104.
Burkhardtsmaier, W., 327.
Burton, C. J., 248.
Burton, E. F., 1, 204.
Busch, H., 207.
von Buzagh, A., 129.
Byerly, P., 57, 59, 60, 67.

Cabannes, J., 381, 387, 404.
Cameron, D. M., 395.
Cameron, G. H., 2, 4, 6.
Campbell, W. B., 182.
Candler, J. H., 336.
Carlson, J. F., 11, 14, 35.
Carman, P. C., 182.
Casimir, H., 363, 369–371, 375.
Cassie, A. B. D., 142, 144, 145, 147–151, 155, 158, 159, 167–169, 182.
Cavaney, B., 193–195.
Chamberlain, N. H., 144, 186.
Chambers, L. A., 247.
Chapman, F. W., 349.
Chauvenet, W., 282.
Chédin, J., 398.
Cheng, H. S., 404.
Chesters, J. H., 114.
Chirnside, R. C., 90.
Chotkevitch, W. J., 362.
Christopher, J., 242, 246.
Clark, G. L., 98, 105, 113, 114, 248.
Clarke, E. T., 12.
Claussen, G. E., 93.
Clay, J., 6, 8, 36, 39, 41, 44, 45.
Clay, P. H., 44.
Clay, R. E., 95.
Clerk, Lt.-Col. H., 280.
Cleveland, F. F., 381, 405.

* Not referred to in text.

Clock, R. C., 248.
Coblentz, W. W., 384.
Cojan, J., 415*.
Coker, E. G., 301.
Collens, H., 320–322, 328, 331, 333, 335, 340, 342, 349.
Compton, A. H., 6, 9, 20, 24, 25, 28, 45.
Conrad, V., 57, 61, 76.
Conrad-Billroth, H. C., 382.
Content, E. J., 121.
Conwell, R. N., 338.
Cook, M., 115, 116.
Cooke, H. L., 1.
Corlin, A., 6.
Cowie, D. B., 33.
Cox, J. H., 336, 338.
Cozzens, B., 339.
Craggs, J. D., 330, 335.
Craib, J., 344.
Crawford, B. L., Jr., 402.
Crary, A. P., 80.
Crosswhite, H. M., 381.
Cruickshank, G., 352.
Czerny, M., 384.

Danforth, W. E., 12, 14, 17, 29.
Daniells, H. E., 180–184.
Daniels, H. E., 151.
Darwin, C. G., 88.
Darwin, Sir H., 309.
Daubenton, 178.
Dauncey, L. H., 90.
Daunt, J. G., 361, 363, 365, 371, 373–375.
Davey, W. P., 103.
Davis, H. T., 196, 197.
Davis, R., 339.
Davison, C., 77.
Debye, P., 91.
Dehlinger, U., 111.
Delatinsky, B., 45.
Delbruck, M., 247.
Delwaulle, M. L., 396.
Demmelmair, A., 10.
Denham, W. S., 146.
Dennis, W. M., 300.
Dennison, D. M., 385.
Derjaguin, B., 136.
Désirant, M., 364.
Dey, L. J. L., 44.
Dieke, G. H., 381.
Dietz, O., 145.
Dippel, C. J., 136.
Dirac, P. A. M., 262, 268–270.
Donnay, J. D., 94.
Dosse, J., 205, 208–211, 217, 218, 237.
Draper, C. S., 294.
Driga, A., 36.
Drysdale, C. V., 294.
Duerden, J. E., 185.
Dufay, J., 381.
Duffendack, O. S., 248.
Dufton, A. F., 126.
DuMond, J. W. M., 95.

Earp, K. E., 189.
Eckart, C., 5.
Edgell, W. F., 404.
Edwards, O. S., 94.
Ehmert, A., 35.
Ehrenfest, P., Jr., 34, 36.
Ehret, L., 125.
Eitel, W., 248.
Elder, J. S., 335, 350–352.
Emmett, P. H., 156–159.
Endell, K., 248.
Euler, H., 34, 258.
Evans, E. A., 334.
Evershed, S., 296.
Ewing, M., 80.
Ewles, J., 182.

Fankuchen, I., 91, 94, 135.
Fechner, G. T., 179.
Fermi, E., 33.
Fielder, F., 338.
Fierz, M., 263, 264.
Filon, L. N. G., 301.
Firestone, F. A., 384.
Fisher, E. A., 147.
Fletcher, H., 122.
Flowers, J. W., 329, 330.
Fogg, G. E., 169.
Foley, H. M., 385.
Forbes, G., 144.
Forbush, S. E., 10.
Forró, M., 45.
Fortescue, C. L., 338.
Foster, G. A. R., 183, 193–196, 198, 199.
Foust, C. M., 336, 337.
Fowler, J. L., 182.
Fox, J. J., 88.
Frampton, L. V., 247.
François, F., 396.
Frankel, J. J., 104.
Franz, E., 181, 182, 248.
Franz, H., 360.
Frenkel, J., 336.
Frenzel, W., 184.
Freon, A., 34, 36.
Freundlich, H., 135.
Frevel, L. K., 105–107.
Fricke, R., 98.
Friedrich, E., 360, 365.
Friedrich-Freska, H., 249*.
Friess, H., 248.
Fritsch, V., 327.
Frocht, M. M., 311*.
Frommer, L., 113.
Froning, A., 193.
Furry, W. H., 11.
Fussell, L., 35.

Galeotti, G., 129.
Galkin, A., 365.
Gallaway, W. S., 395.
Gard, S., 247.
Gardner, G. F., 337.
Garvie, W. M., 161.
Gaskell, J., 45.
Gaskell, T. F., 80.
Geiger, H., 35.
Geisler, A. H., 241, 248,
van Gemert, A., 36, 44, 45.
Gerschlauer, W., 362.
Getting, I. A., 24.
Gibson, J., 109.
Gilbert, C. W., 35.
Gildart, L., 248.
Gill, P. S., 35.
Ginsburg, N., 385.
Ginsburg, V., 371.
Glaser, W., 205, 207, 208, 210, 219, 225, 242.
Glocker, G., 396, 402, 404.
Gockel, A., 1.
Gold, E., 3.
Golde, R. H., 327, 330–334, 340–342, 345, 349.
Goobar, S. W., 361, 373.
Goodings, A. C., 193.
Goodlet, B. L., 327, 336.
Gore, G., 283.
Gorter, C. J., 363.
Goss, G., 298.
Goss, N. P., 116.
Gott, J. P., 317.
Goubeau, J., 402, 415.
Graaf, G. N., 44.
de Graaf, J. E., 95.
de Gramont, A., 309
Grant, C. Kerr, 80.
Grant, K., 6.
Graves, A. C., 36.
Graziadei, H., 45, 46.
Green, L., 121.
Greenhill, A. G., 182.
Gregory, J., 183, 193–195.
Greisen, K., 23, 42–44.
Greninger, A. B., 89.
Grimm, H. H., 381.
Grivet-Meyer, T., 36.
Groenewold, K., 369.
Gross, B., 5, 17.
Gross, I. W., 337.
Grunewald, H., 337.
Guillaume, C.-E., 280.
Guinier, A., 101, 102.
Funn, R., 25.
Günther, F., 232.
Gutenberg, B., 55–57, 68.

Haagen, E., 226,
de Haas, W. J., 362, 368, 371, 372.
Hagenguth, J. H., 338, 339.
Hale, J. B., 395, 413.
Hall, C. E., 248.
Hall, D. B., 258.
Halliday, E. C., 334.
Hamaker, H. C., 136.
Hamilton, D. T., 405.
Hamilton, J., 23, 24, 258.
Hanawalt, J. D., 105.
Hanov, B., 327.
Hansson, E., 337.
Harding, S. W., 103.
Hardy, J. D., 383, 385.
Hardy, J. I., 188.
Harish-Chandra, 261,
Harker, D., 240, 241, 248.
Harland, W. B., 80.
Hass, G., 238, 249.
Hauser, E. A., 248.
Hawley, G. G., 204.
Haworth, F. E., 101.
Hearson, T. A., 272.
Hebb, M. H., 22.
Heck, N. H., 77, 80.
Hedges, J. J., 146, 156.

* Not referred to in text.

Heidenreich, R. D., 223, 224, 241.
Heinmets, F., 246.
Heisenberg, W., 22, 34, 258.
Heitler, W., 11, 22, 34, 37, 258, 270.
Heller, G., 367.
Henderson, J. T., 337.
Hendricks, S. B., 99.
Henle, W., 247.
Henning, H. J., 202*.
Henry, P. S. H., 151.
Hercus, E. O., 143.
Herd, J. F., 349, 350.
Herschel, J. F. W., 284.
Hertel, K. L., 182, 184.
Hertz, H., 306.
Herzog, G., 32.
Hess, V. F., 1, 2, 10, 36, 45, 46.
Hetényi, M., 301.
Hibben, J. H., 415.
Hickman, J. W., 366.
Hicks, B., 95.
Hilbery, N., 33, 39.
Hill, A. V., 147.
Hill, S., 122.
Hillier, J., 220, 223, 225, 226, 229, 230, 232, 233, 236, 237, 243–245.
von Hippel, A., 336.
Hirschlaff, E., 373.
Hoag, J. B., 33, 39.
Hobson, Ltd., 300.
Hodges, D. B., 320, 335, 349–352.
Hoerlin, H., 34.
Hoffert, H. H.., 318.
Hoffmann, G., 35.
Hofmann, U., 248.
Hoghton, H. P., 299.
Holleman, L. W. J., 135.
Holtzapffel, C., 303.
Holtzapffel, J. J., 303.
Holzer, R. F., 320, 336, 348.
't Hooft, C. G., 44.
Hopf, M., 385.
Horseman, A., 371, 374.
Huberty, A., 184.
Hudspeth, E. J., 30.
Huggins, L. G., 336.
Hull, A. W., 91.
Hultgren, R., 103.
Hume-Rothery, W., 107, 116.
Humphreys, W. J., 336, 339.
Husemann, E., 248.
Hutter, R. G. E., 213, 214, 218, 219, 242.
Hutton, C. O., 91.

Ichimiya, T., 33.
Imhoff, C. E., 104.
Ingold, G. K., 395, 413.
International Wool Textile Organization, 182.
Ising, G., 276.

Jackson, L. C., 358.
Jakob, A., 246.
Jánossy, L., 34.
Jay, A. H., 96, 102, 114.
Jeffreys, H., 53, 57, 58, 62–68, 70, 72–78, 80–82, 281.
Jenkins, H. G., 108.
Jesse, W. P., 19–22, 30, 34, 39.
Jessop, G., 167.
Jobling, H. B., 338.
Johnson, T. H., 8–10, 12, 14, 17, 19, 20, 26, 30, 33.
Jolley, A. C., 294.
Jones, F. W., 109.
Jones, H., 32.
de Jong, H. G. Bungenberg, 131, 134, 135.
de Jong, H. G. Bunbenberg, and collaborators, 137–140.
Justi, E., 359, 360, 362, 365, 372.

Kahovec, L., 406.
Kan, L., 365, 366.
Karweil, J., 402.
Kastler, A., 380.
Katz, D. L., 248.
Katz, J. R., 156.
Kausche, G. A., 247.
Keeley, T. C., 373.
Keesom, W. H., 363, 364.
Kehler, H., 239.
Keller, F., 241, 248.
Kelvin, Lord, 290.
Kemmer, N., 261.
Kikoin, I. K., 361, 373.
King, A. T., 142, 152, 161–165, 181.
King, E. J., 104.
King, G., 147, 148.
Kip, A. F., 327.
Kirkpatrick, P., 88.
Klein, O., 4.
Klemperer, O., 204, 213, 226, 246.
Klit, A., 414.
Knaysi, G., 246.
Kobayasi, M., 34.
Koch, H. W., 248.
Koch, K. M., 372.
Kochendorfer, A., 111.
Koets, P., 135, 136.
Kohl, W. H., 204.
Kohler, R., 129.
Kohlrausch, K. W. F., 382, 389, 396, 400–402, 404, 406, 409, 414, 415.
Kolhörster, W., 2, 6, 45.
Köppl, F., 402.
Korff, S. A., 12, 20, 37.
Kossel, W., 223, 224.
Kosten, C. W., 122.
Kozeny, I. J., 182.
Kramer, J., 360, 365.
Kratky, O., 101.
Kruyt, H. R., 131, 135, 136.
Kuehni, H. P., 337, 338.
Kumpf, K., 416*.

Laby, T. H., 352.
Ladd, W. A., 248.
van Laer, P. H., 363, 369.
Lamb, H., 65, 76, 306.
Lammel, E., 219.
Landau, L., 367.
Langmuir, I., 132, 136.
Langseth, A., 403, 405, 414.
Larsen, A., 319, 333.
Lasarew, B., 365, 366.
Laubengayer, A. W., 103.
v. Laue, M., 372, 373.
Lauffer, M. A., 135, 247.
Lear, C. M., 327.
LeBeau, D. S., 248.
Leckie, A. H., 395, 413.
Lecomte, J., 384, 398, 404.
Lee, E., 395.
Lee, E. S., 336, 338.
Lees, C. H., 143.
Legg, J. W., 336.
Lehmann, I., 61, 67, 68.
Lemaitre, G., 7.
Lembke, A., 242, 246.
Lenard, P., 355*.
Lennard-Jones, J. E., 163.
Leprince-Ringuet, L., 8.
Lewis, W. W., 336, 337.
Lindholm, F., 35.
Linnet, J. W., 402.
Lippert, G. D., 337.
Lipson, H., 94, 96–98, 111.
Lloyd, E. H., 113,
Lloyd, Ll. S., 126.
Lloyd, W. L., 338.
Locher, G. L., 12, 14, 17, 29, 32.
Loeb, L. B., 325, 327.
London, F., 370, 371, 375.
London, H., 366, 367, 370, 373, 375.
Lonsdale, K., 88, 91.
Loring, H. S., 247.
Ludloff, H. F., 124.
Lukesh, J. S., 98.
Luria, S. E., 247.
Lutkin, F. E., 335, 349, 350, 352.

Mas, D. Y., 121.
Maass, O., 182.
MacArthur, I., 95.
McAuley, P. H., 336.
Mc Bain, J. W., 129, 241, 248.
McCann, G. D., 320, 323, 327, 330, 331, 334, 337, 339, 340, 342, 343.
McCune, F. K., 298.
MacCurdy, E., 311.
McEachron, K. B., 319, 320, 323, 327, 333, 338–342.
Macelwane, J. B., 55.
McKeag, A. H., 108.
McKinley, G. M., 236.
McLane, G. L., 327.
McMorris, W. A., 338.
McNeill, J. J., 352.
Mahl, H., 232, 246, 248.
Maier-Leibnitz, H., 34.
Malan, D. J., 320–322, 328, 331, 333, 335, 340.
Marsh, M. C., 144, 187, 189.
Martin, A. E., 88.
Martin, A. J. P., 82, 186, 187.
Martin, D., 248.
Martindale, J. G., 175, 193, 195, 197, 198, 200, 201.

* Not referred to in text.

Marton, L., 204, 213, 214, 220, 226, 228, 229, 234, 236, 238, 239, 242, 243, 247, 248.
Mason, C. A., 120, 122.
Mason, C. C., 309.
Massignon, D., 398.
Matthew, J. A., 144, 193.
Matthews, J. M., 142.
Matthias, A., 327, 333.
Matthieu, J. P., 398, 413.
Mauguin, Ch., 390.
Maxwell, J. C., 278, 279, 305.
Maze, R., 34, 36.
Mazeleo, L. Y., 98.
Mecke, R., 416*.
Mecklenburg, W., 237.
Meek, J. M., 325–327, 330, 335.
Megaw, H., 111.
Meissner, W., 358, 360, 363, 364.
Mellor, J. W., 282.
Melnick, J. L., 247.
Mendelssohn, K., 358, 361, 363–366, 368–371, 373–375.
Menkart, J., 186.
Menzies, A. C., 383.
Merrill, F. H., 336.
Mesnager, A., 299.
Meyer, E., 124.
Mikura, Z., 372.
Milatz, J. M. W., 276.
Milligan, W. O., 103.
Millikan, R. A., 2, 4, 6, 12, 26.
Mills, H. R., 383.
Miranda, F. J., 328.
Misener, A. D., 366, 370, 375.
Moir, J., 120, 122.
Moktar, M., 127.
Møller, C., 38, 257, 258.
Mollet, P., 384.
Moncrieff, R. W., 274.
Montgomery, C. G., 15, 30, 33, 35, 37.
Montgomery, D. D., 15, 30, 33, 35, 37.
Moore, J. R., 358, 363, 366.
Morgan, D. E., 334.
Morgan, J. E., 34, 36.
Morgan, K. Z., 34–36.
Mosharrafa, A. M., 127.
Mott-Smith, L. M., 31, 32.
Mrgudich, J. N., 248.
Mudd, S., 239, 242, 246, 247.
Müller, A., 95.
Müller, E., 183, 184.
Müller, H. O., 229, 237, 240, 248.
Murphy, G. M., 389.
Murphy, M. J., 240, 241.

Nagelschmidt, G., 99, 104.
Nakhutin, J. E., 362, 366, 369.
Neale, S. M., 146, 161.
Neddermeyer, S. H., 16, 32, 33, 35.
Neher, H. V., 2, 8, 12, 26.
Nelson, J. B., 97.
Neugebauer, T., 336, 372.
Nicholls, F. G., 352.
Nichols, J. E., 202*.
Nickson, A. F. B., 352.
Nielsen, H. H., 395.
Nielsen, J. R., 416*.
Nielsen, L. E., 395.
Nielsen, W. M., 30, 34–36.
Nishina, Y., 4, 33.
Nissen, H. F., 244.
Niven, C. D., 144.
Nordheim, G., 34.
Nordheim, L. W., 11, 22, 34.
Norinder, H., 338, 349, 350.
Noto, H., 356*.

O'Brien, H. C., 236.
Occhialini, G. P. S., 35.
Ochsenfeld, R., 363.
Okayama, T., 34.
Oldham, R. D., 55.
Onnes, H. Kamerlingh, 360, 362, 363, 365, 366.
Oosterkamp, W. J., 95.
Oppenheimer, J. R., 11, 14, 35, 38.
Ornstein, L. S., 276.
Oshry, H. I., 105.
Ostwald, W., 129.
Otis, R. M., 2.
Owen, E. A., 100.

Partridge, J. H., 110.
Pascoe, K. J., 111.
Pauli, W., 263, 264, 268, 270.
Pauling, L., 160, 409.
Pawlek, F., 248.
Peck, V. G., 241.
Peierls, R., 16, 22, 32–34.
Peirce, F. T., 142, 158, 159.
Pendzich, A., 232.
Peng, H. W., 22, 34, 37, 258, 270.
Perry, F. R., 331, 337, 338, 346.
Petch, N. J., 98.
Peters, J. F., 336.
Pfister, R. J., 381.
Pforte, F., 35.
Pfotzen, G., 12, 14, 23, 30.
Phillips, A. C., 307.
Phillips, H., 161.
Phillips, W. E., 335, 350–352.
Picard, R. G., 248.
Picht, J., 204.
Pickarski, G., 416*.
Pickering, W. H., 8, 26.
Pickup, L., 100.
Pirie, N. W., 135.
Pittman, R. R., 338.
Placzek, G., 37, 387, 389
Plett, G., 61, 66.
Plummer, C. H., 195.
Pohlmann, R., 124.
Pollard, A. F. C., 275, 280, 289, 297, 298, 301, 303, 307–310.
Pomerantz, M. A., 33.
Pontius, R. B., 369, 370.
Pool, B., 146, 147.
Porter, J. S., 115.
Poynting, J. H., 290, 291
Prebus, A. F., 204, 232.
Preston, T., 148.
Price, A. L., 338.
Prieboch, J. A., 35, 36.
Proca, A., 260.

Quimby, S. L., 128*.

Rademakers, A., 371, 372.
Raisin, C. G., 395, 413.
Ramberg, E. G., 241, 244, 245.
Ramo, S., 217, 231, 338.
Ramsauer, C., 204, 218.
Ramsey, W. E., 15, 30, 33–36.
Randall, H. M., 385.
Rank, D. H., 381.
Rannie, J. L., 300.
Ransley, C. E., 116.
Rao, B. S. Madhava, 271, 372.
Rasetti, F., 33.
Rayleigh, Lord, 276.
Read, D. N., 8.
Rebsch, R., 230, 243.
Recknagel, A., 204, 217, 218, 237.
Redlich, O., 394, 395.
Rees, W. H., 144.
Rees, W. J., 107.
Regener, E., 2, 12, 34.
Reitz, A. W., 382.
Renfrew, M. M., 402.
Reuleaux, F., 272. 273,
Reynolds, D. H., 105.
Richard, J., 193.
Richards, A. G., 248.
Richards, T. Ll., 115.
Richter, C. F., 55–57, 68, 78.
Riley, D. P., 97, 102, 105, 109.
Rivlin, R. S., 86.
Rjabinin, J. N., 362.
Roberts, G. E., 33.
Robertson, J. M., 102.
Robertson, R., 88.
Robinson, G. D., 314, 315, 317, 318.
Robley, 34, 36.
Rochester, G. D., 34.
Roll, A., 125.
Rolt, F. H., 280, 299, 302, 309.
Ronnebeck, H. R., 102.
Rooksby, H. P., 90, 91, 103, 104, 106–108, 110, 111, 114, 116.
Rose, M. E., 30.
Rosenfeld, L., 257.
Rosental, S., 258.
Rosenthal, J. E., 389.
Rossi, B., 6, 9, 23, 32–35, 39, 258.
Rousset, A., 380, 381.
Rowett, F. E., 293.
Rubner, N., 143.
Rusher, M. A., 338.
Ruska, E., 204, 213, 215, 218, 220, 222, 226, 232, 238, 240–243, 246, 247.
Ruska, H., 416*.
Ruthemann, G., 226.
Rutherford, E., 1.
Rutherford, H. M., 80.

* Not referred to in text.

Sabata, E. R., 26.
Sabine, P. E., 123.
Sage, C. G., 402.
Sakata, S., 34.
Sass, S., 204.
Sauter, F., 367.
Saxl, I. J., 193, 194.
Schaefer, V. J., 240.
Schäfer, K., 402.
Schein, M., 19–22, 30, 32, 34, 35, 39.
Scherrer, P., 91, 109, 111.
Schiebold, E., 232, 248.
Schiff, L. J., 221, 225, 228.
Schindler, H., 34.
Schlink, F. J., 284, 287–291.
Schloesing, T., 145.
Schmeisser, K., 34, 36.
Schmerwitz, G., 58.
Schmidt. G., 125.
Schmidt, R. W., 242.
Schmitt, F. O., 248.
Schoen, A. L., 248.
Schoeneck, H., 363.
Schoeneich, H., 278.
Schonland, B. F. J., 45, 320–322, 324–327, 329–331, 333–335, 340, 344, 345, 349–352.
Schoon, Th., 248.
Schrenk, H. H., 105.
Schuhmeister, J., 143.
Schultz, E. W., 247.
Schusterius, C., 248.
Schwegler, A., 36.
Schwerdt, C. E., 247.
Scrase, F. J., 56, 65, 66, 314, 316–318, 334, 347, 348.
Sears, J. E., 309.
Sears, W. C., 395.
Seka, R., 416*.
Semmler-Alter, E., 248.
Shalnikov, A., 368, 369.
Shaw, B. T., 248.
Shaw, H., 300.
Shaw, W. F., 300.
Shepelev, J. D., 362.
Sherman, J., 409.
Shipley, J. F., 333.
Shoenberg, D., 358, 359, 363, 368, 370, 371, 375.
Shorter, S. A., 146, 187–189.
Shubnikov, L., 362.
Shuck, O. H., 126.
Shutt, R. P., 32, 33.
Sieberg, A., 78.
Siegbahn, K., 213, 231.
Siegel, S., 124.
Silsbee, F. B., 362.
Simon, A., 415*.
Simpson, G. C., 314–319, 334.
Sinclair, H., 96.
Sitaraman, W L., 127.
Sizoo, G. J., 365, 366.
Skehan, J. W., 93.
Sliepcevich, C. M., 248.
Smeloff, N. N., 338.
Smith, C. S., 111.
Smith, F. R., 103.
Smith, S. L., 113.
Snoddy, L. B., 320, 322.
Snyder, R. L., 237.
Sokolov, S., 125.
Sommerfeld, A., 374.
Speakman, J. B., 144, 146, 154, 155, 159, 186, 188.
Spencer-Smith, J. L., 199, 200.
Sporn, P., 338.
Stanbury, G. R., 180–182, 184, 193.
Stanley, W. M., 136, 247, 248.
Stark, J., 372.
Starks, H. J. H., 204.
Starling, S. G., 283.
Steiner, K., 362, 363.
Steinke, E., 34.
Steinmaurer, R., 10, 45.
Stekolnikov, I. S., 320, 327, 338–340.
Stevenson, E. C., 16, 32.
Stickley, E. E., 111.
Stitt, F., 401, 402.
Stockmair, W., 416*.
Stoermer, C., 2, 7.
Stokes, A. R., 111.
Stoneley, R., 59, 65, 75, 82.
Stott, E., 186.
Stott, V., 298.
Stranski, J., 248.
Street, J. C., 9. 16, 32, 33, 35.
Stringfellow, W, A., 146, 161.
Sturkey, L., 223, 224.
Sullivan, R. R., 182.
Sundberg, K., 306.
Sutherland, D. M., 143.
Sutherland, G. B. B. M., 395.
Svartholm, N., 213.
Swan, E., 158.
Swann, W. F. G., 5, 7, 11, 12, 14, 15, 17, 19–21, 24, 25, 27–30, 33–36, 38, 41–44.
Sykes, C., 107.
Taketani, M., 34.
Takeuchi, M., 33.
Talalay, P., 248.
Tatel, H., 103.
Taylor, A., 96, 98.
Taylor, Ltd., 300.
Teller, E., 156–159, 394.
Teunissen-van Zijp, L., 135.
Thomas, D. E., 90, 112, 113.
Thomason, J. L., 339.
Thomassen, L., 240, 247.
Thompson, J. L., 45.
Thomson, J. J., 38, 290, 291.
Thomson, J. W., 395, 413.
Thornton, W. M., 336.
Tillotson, E., 82.
Tiselius, A., 247.
Tizza, L., 403.
Todd, H. A. C., 199, 200.
Tomkins, J. A., 309.
Tordella, J. P., 248.
Torok, J. J., 338.
Townend, S., 180, 184.
Trost, A., 128*.
Trouton, F. T., 146, 147.
Trumpy, 396.
Turkevich, J., 248.
Turner, H. H., 60, 64, 82.
Turner, L. B., 283.
Turner, R. N., 45.
Tuve, M. A., 6.
Tuyn, W., 360, 362.
Urquhart, A. R., 146, 158.
Valeev, C., 338–340.
Vallarta, M. S., 7.
Vance, W. A., 228–230.
Viviani, E., 193.
Volckmann, J. E., 122.
Vold, M J., 129.
Vold, R. D., 129, 248.
Voogd, J., 362.
de Vos, P. J. G., 36.
Vose, R. W., 195.
Wadati, K., 65.
Wagner, C. F., 327, 337, 340, 342, 343.
Wagner, J., 400, 401, 406.
Waldorf, S. K., 337.
Walker, Sir G., 196, 198.
Walker, H. N., 36.
Walter, B., 319, 322, 331.
Ward, N. E., 416*.
Wcrren, B. E., 101, 109.
Watson, B. B., 95.
Watson Watt, R. A., 349, 350.
Weber, L. R., 385.
Webster, G. H., 337. 346.
Weiser, H. B., 103.
Weisz, P. B., 30, 32.
Weisz, R. S., 103.
Welker, H., 373, 374.
Wenzel, R. N., 167.
Wergeland, H., 34.
Westerhoff, H., 360.
Weygand, C., 248.
Whipple, F. J., W. 149, 334, 345, 347, 348.
Whipple, R. S., 309.
White, W. B. L., 126.
Whitehead, W. and J., Ltd., 193.
Wiegand, W. B., 248.
Wierama, J. T., 8, 34, 36.
Wiggins, E. J., 182.
Wigner, E. 389.
Wilde, W. T., 107.
Wildman, A. B., 203*.
Wilkinson, H., 184.
Williams, A. M., 146.
Williams, E. J., 33, 38.
Williams, R. C., 239, 240.
Williamson, C., 127.
Wilson, A. H., 270.
Wilson, A. J. C., 96, 97.
Wilson, C. L., 395, 413.
Wilson, C. T. R., 317, 343, 344, 347, 348.
Wilson, E. B., 394, 395.
Wilson, J. T., 57.
Wilson, V. C., 22, 32, 34, 35, 44, 45.
Winson, C. G., 179, 181, 182, 185.
Wittek, H., 400, 404
Wolf, F., 335.

* Not referred to in text.

Wollan, E. O., 19–22, 30, 34, 39.
Womersley, J. R., 183, 193.
Wood, W. A., 110, 113, 116.
Woods, H. J., 188.
Woodward, R. H., 30, 33, 35, 36.
Woog, P., 169.
Wooster, N., 87.
Wooster, W. A., 85.
Workman, E. J., 320, 322, 336, 348.
Wormell, T. W., 335, 343–345, 347, 348.
Wright, N., 385.
Wrightson, T., 280.
Wu, C. K., 395.
Wu, T. Y., 404.
Wyckoff, R. W. G., 239, 240.
van Wyk, J. W., 335, 350–352.
Wyllie, M. R. J., 116.

Yates, D. E., 104.
Yavorsky, V. V., 327.
Young, R. T., Jr., 35.
Young, R. W., 126.
Ypsilanti, G. P., 404.
Yukawa, H., 32, 34.
Yule, G. U., 196–198.

Zaduk, H., 337.
Zalesaki, A. M., 327.
Zanstra, H., 20, 26.
Zeiller, O., 35.
Zickner, G., 372.
Ziegler, W. T., 361, 366.
Zocher, H., 135.
Zöppritz, K., 54,.
Zwikker, C., 156.
Zwikker, C. J., 123.
Zworykin, V. K., 229, 230, 232, 233, 237, 241.

COSMIC RAYS

By W. F. G. SWANN,
Bartol Research Foundation of the Franklin Institute, Swarthmore, Pennsylvania

§1. INTRODUCTION

General plan of the Report

This *Report* will be divided into two parts. Part I will constitute a running account of the development of the subject, with but brief references to the earliest work. Part II will contain an elaboration of certain portions of Part I, and will deal with matters of detail and critical discussion.

PART I. SURVEY OF THE COSMIC-RAY PROBLEM

§2. EARLIER WORK

The dawn of cosmic-ray research started when McLennan and Burton (1903) and Rutherford and Cooke (1903) found that the ionization which occurred in a closed vessel could be reduced by surrounding the vessel with thick shields. There is radium emanation in the air, but only about 250 gm. of it in the whole of the earth's atmosphere. Since this small density of radioactive materials produced in the air five times as much ionization as is now known to be produced by the cosmic radiation at sea level, and since the soil itself contains an appreciable amount of radioactive contamination, as indeed do all ordinary substances, it will be readily seen that, in an epoch in which these matters were gradually coming to light, there was reason for considerable doubt as to whether the residual ionization in a supposedly uncontaminated vessel was really representative of anything but "residual contamination", in the vessel or in the immediate surroundings. Today, indeed, we know that most of it was attributable to such causes. The disentanglement of the contaminative effect from the more fundamental phenomenon commenced when investigators sought to measure the variation of residual ionization with altitude. Gockel (1910, 11) made three balloon ascents to 2500, 2800 and 4000 m. respectively, and found a slight decrease of the ionization with altitude, but not as much as could be expected if the bulk of the gamma radiation comes from the ground. From Gockel's observations even a small increase of ionization with altitude can be derived if his values are reduced to normal pressure to allow for the fact that his chambers were not airtight. However, his observations were influenced by instrumental defects, and showed such great fluctuations that it was not possible to draw any definite conclusions from them. Hess (1911, 12a) made ten balloon ascents up to 5400 m. with much improved instruments, and found that after a small decrease from sea level to about 1000 m., the ionization began to increase very decidedly, and at 5000 m. reached values several times as great as at sea level. Hess (1912b) concluded that his observations could be explained only by assuming the existence of a new radiation of extra-terrestrial origin and of a penetrating power exceeding

many times the penetrating power of gamma rays. His work was recognized by the Nobel Prize.

Hess's observations were followed up to an altitude of 9000 m. by Kolhörster (1913, 14), where the ionization reached the magnitude of 80 ions per cc. per second as compared with from 4 to 10 ions per c.c. per second, which represented about the value customary for observers of those times to find at sea level.

The observations of Hess and Kolhörster thus pointed to the existence of a radiation of extra-terrestrial origin, subsequently to be known as the *penetrating radiation*, and later as the *cosmic radiation*. These observations, interpreted in terms of an exponential law of mass absorption, led to an absorption coefficient of about 0·57 per metre of water, which represents about 10 times the penetrating power of the most penetrating gamma rays then known.

§3. EXPERIMENTS AND CONCEPTS OF THE INTERMEDIATE EPOCH

(i) *Cosmic-ray intensities and absorption.*

Observations on variation of the intensity of the radiation with altitude received a new impetus in the work of Millikan and his collaborators which appeared in several papers commencing in 1923. Millikan and Bowen (1923, 26) and Regener (1932) sent to high altitudes sounding balloons equipped with photographically recording electroscopes. Experiments on mountain peaks were carried out by Millikan and Otis (1926), and Millikan and Cameron (1926) carried out investigations in snow-fed lakes (snow-fed to minimize radioactive contamination) at high altitudes, while Bowen, Millikan and Neher (1933) carried out observations in airplanes. Naturally, the object of these experiments was to provide the basis for going as far as possible towards a decision as to the nature of the radiation. However, at the time when these experiments were made, physics was in no very satisfactory condition to account for the degree of penetration implied. Beta-ray electrons, capable of traversing a millimetre or so of lead, represented the most penetrating electrically charged particles known. It is true that, according to the views of the day, electrons lost energy only by ionization, so that increase of energy carried with it a proportional increase in penetrating power. However, even on such views, now of course outmoded, it was necessary to assume energies comparable with more than a billion electron volts to secure passage of the radiation through the atmosphere. Today we do not squirm at the assumption of energies a million or more times larger. However, in the early days of cosmic-ray research, such energies were regarded as prohibitive, in spite of the fact that in an entirely different field—that of auroral phenomena—Stoermer (1904) had found it necessary to suggest energies comparable with a billion electron volts to account for an explanation of the *Aurora Borealis* as a phenomenon caused by charged particles emanating from the sun.* It seemed more hopeful to turn to some radiation of the gamma-ray nature in order to account for the great penetration ; for although gamma rays

* For any magnetic latitude there is a minimum momentum which must be exceeded by a charged particle if it is to reach the earth's surface at that latitude and not be turned back into space by the earth's magnetic field. The minimum momentum increases with decrease of latitude, and Stoermer's calculations show that in order to account for the *Aurora* at latitudes as low as those at which it is observed, momenta corresponding to energies as high as a billion volts must be assumed in the case of electrons.

from radium are far too soft to function as cosmic rays, they are much more penetrating than the corresponding electrons. Moreover, at the time there was a convenient ignorance of the details of gamma-ray absorption, which enabled us, without too serious a protest from the theories, to speculate as to the possibility of very high penetration by gamma rays, or photons, as we now call them, of sufficiently high energy. The quantum theory had already associated high photon energy with high frequency, and so interest centred for a time upon the frequencies of which the cosmic-ray photons might be supposed to be made up.

One does not perhaps always realize the enormous extrapolation involved in attempting to extend such laws as were known for electrons and photons of ordinary nuclear energies to those which were anticipated for cosmic-ray energies. Thus, for even moderate cosmic-ray energies, the extrapolation was of the order of magnitude of 10,000 which, as applied to the forms of the laws invoked, would represent the equivalent of a physicist drawing a curve on a piece of paper one foot square covering the range in which his laws had been tested, and then extrapolating for the distance of two miles for the purpose of drawing conclusions outside of the range. In some cases, the extrapolation would be as much as 200 miles or more. However, even apart from uncertainty as to the laws themselves, the problem was beset with sub-problems of a complex and diffuse nature. The ideal goal, in terms of the beliefs of the times, was to assign incoming photons with an arbitrary spectral distribution in energy and then deduce the nature of this distribution by tracing its consequences in the facts as regards the depth ionization curve in the atmosphere and other media, such as lakes, etc.

The simplest case was that represented by an incoming radiation with a single mass absorption coefficient μ, the intensity I_0, at a place where the pressure is negligible, being the same in all directions and representing the number of photons received per unit solid angle per second per square centimetre (with plane perpendicular to the axis of the solid angle). If we consider a point at a depth H below the top of the homogeneous atmosphere,* and draw from it an elementary solid angle $d\omega = \sin\theta\, d\theta\, d\phi$ defined by the element $d\theta$ in zenith angle and $d\phi$ in azimuth, the number of photons passing per second per square centimetre (perpendicular to the axis of the solid angle) is dN, where

$$dN = I_0 e^{-\mu H \sec\theta} \sin\theta\, d\theta\, d\phi. \qquad \ldots\ldots (1)$$

If we should assume that, as a result of a photon travelling through an ionization chamber, there is a probability of ionization proportional to the distance traversed in the chamber, either as a result of direct action or as a result of secondary emission from the walls, so that β is the average number of ions produced per centimetre of path by a ray within the vessel by direct or indirect action, then it is easy to show† that the number of ions produced per second within the ionization chamber of volume V is

$$N = 2\pi\beta V I^0 \int_1^\infty \frac{e^{-\mu H x}}{x^2}\, dx \qquad \ldots\ldots (2)$$

where the integral is known as Gold's (1909) integral. Even here we do not have a true exponential in H, but the custom in earlier work was to adjust the value of

* The homogeneous atmosphere is regarded as obtained by compressing the actual atmosphere until at each point it attains a density equal to that of water.

† Consider a tubular section of the vessel (figure 1) with axis OP, length l, and cross-section da. The ionization produced per second in the cylinder by the rays of intensity I_ω travelling within the solid angle $d\omega$, whose axis is OP, is $\beta I_\omega l\, da\, d\omega$. If we divide the complete volume into tubes

[*Continued on page* 4

μ to fit the experimental results as well as possible, and even to attempt to define a set of three or four different primary μ's which, operating independently through the procedure cited, would give the best fit to the experimental curves. The desire was to associate with each of these values of μ an incoming radiation of a definite photon frequency such as to give the value of μ in question,* and relate the energies associated with the photon frequencies in question to the energies of certain atom-building processes, as suggested by the so-called mass defects of the atoms as revealed by Aston's (1927) measurements of the masses of the isotopes. Millikan and Cameron (1928, 31) felt that their observations gave evidence of the existence of four bands in the spectral distribution which they sought to

of the above kind, all parallel to OP, and sum, we obtain for the whole ionization due to the radiation coming within equivalent solid angles $d\omega$, the value

$$dN\gamma = \beta V I_\omega d\omega,$$

where V is the volume of the vessel.

Writing $d\omega = \sin\theta\, d\theta\, d\phi$, where θ is zenith angle and ϕ is azimuth, and using (1), we readily obtain (2) on changing the variable from θ to x ($=\sec\theta$).

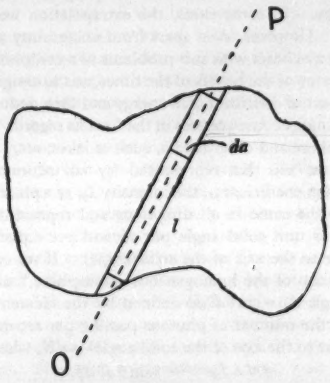


Figure 1.

* This was first attempted by Millikan and Cameron (1926) on the basis of the Compton-Ahmad formula for the absorption of gamma rays. Later (1931) they used the Klein-Nishina (1928) formula, which gives a definite absorption coefficient for a photon frequency ν after the primary photons have penetrated so far into the atmosphere that they have attained equilibrium with the secondary and higher order Compton photons to which they give rise and with the accompanying electrons. The interpretation becomes complicated to the extent that the aforesaid equilibrium is not realized at the places of observation, and it is further complicated when, in the case of the higher frequencies, some of the Compton electrons compare (according to the theories of the day) in penetrating power with the penetrating power of the photons themselves. Of course, all of these considerations become modified in the light of our present knowledge concerning "pair-production" to be discussed later.

The Klein-Nishina formula is as follows:

$$\mu = \frac{2\pi Ne^4}{m^2c^4}\left\{\frac{(1+\alpha)}{\alpha^3}\left[\frac{2(1+\alpha)}{1+2\alpha} - \frac{1}{\alpha}\log_e(1+2\alpha)\right] + \frac{1}{2\alpha}\log_e(1+2\alpha) - \frac{1+3\alpha}{(1+2\alpha)^2}\right\},$$

where $\alpha = h\nu/mc^2$, ν is the frequency, h is Planck's constant, m is the electronic mass, c is the velocity of light, and N is the number of electrons per cc.

associate with the building of helium, oxygen,* silicon,* and iron,* out of hydrogen, the building process presumably taking place in interstellar space.

Quite apart from matters concerned with the complete change in the basic concepts as regards the behaviour of high-energy photons and electrons, to be discussed later, arguments such as the foregoing are attended with several difficulties. In the first place, there are the difficulties concerned with the interpretation of a single incoming frequency in terms of a pure absorption coefficient, and involving the degree of attainment of equilibrium. Then the analysis of the intensity-altitude curves into a number of pure exponential contributions is a dangerous and uncertain procedure. Eckart (1934), who has made a detailed analysis with as much mathematical rigour as possible, failed to confirm the banded structure suggested by the earlier analyses.

A less ambitious problem than that of actually analysing the incoming radiation into exponential components is that of deducing from ionization-chamber observations, which are concerned with the radiation from all directions, the specific intensity corresponding to a single direction. This problem has been solved by Gross (1933), and his method has been extensively used in interpreting the later observations with ionization chambers.† It is to be observed, however, that the method devised by Gross is only applicable provided that the absorption of the radiation in a given direction depends only upon the water-equivalent path, and is otherwise independent of the actual distance travelled. Such an assumption is invalid in the light of our present knowledge, to be cited later, that some of the radiation disintegrates spontaneously, after a finite life, in analogy with radioactive disintegration. Moreover, it is only by a fortunate mathematical circumstance that, as pointed out by Swann (1935), when the energies concerned are sufficiently high, the secondary and higher-order radiations generated in the atmosphere, either according to classical or later quantum principles, perpetuate the path of the primary entity. Failure of this condition would complicate considerably any analysis such as that of Gross.

While the older ionization chamber data were in the process of study, with a view to disentanglement of the competing considerations, a new flock of evidence, experimental and theoretical, came upon the scene and changed very considerably the fundamental pictures which formed the basis of our thinking in cosmic rays. Fortunately, the Geiger-Mueller counter, a tool more discriminating than the ionization chamber, particularly as regards directional effects, began to develop as an instrument of precision and served to elucidate many matters whose eluci-

* Or elements close to these elements in mass defect.

† If we suppose that outside the atmosphere the radiation is of equal intensity in all directions in a hemisphere, and that $f(s)$ is the intensity of the vertical radiation after penetrating to a water equivalent depth s in the atmosphere, then the total ionization in a vessel of volume V at the depth s is determined, in a manner analogous to that used in the deduction of equation (2), and is $2\pi\beta V I_s$, where, if $\sigma \equiv s/\cos\theta$,

$$I_s = \int_0^{\frac{\pi}{2}} f(s/\cos\theta)\sin\theta\, d\theta = s\int_s^{\infty} [f(\sigma)/\sigma^3]d\sigma. \qquad \ldots\ldots (3)$$

Gross has shown that this equation may be solved for $f(s)$ by the relation

$$f(s) = I_s - s(dI_s/ds) = I_s\left[1 - \left\{d(\log I_s)/d(\log s)\right\}\right]. \qquad \ldots\ldots (4)$$

dation would have been impossible with ionization chambers alone. In particular, a Geiger-counter telescope* permits the realization, in measurable form, of the *specific intensity* for any direction, a quantity already referred to and defined as the *number of rays received per unit solid angle per second per unit of area drawn perpendicular to the solid angle.*

(ii) *The discovery of the latitude effect and its bearing upon cosmic-ray theory*

Two decades ago, the principal arguments against the assumption of charged particles as cosmic rays were : (1) the difficulty of accounting for the degree of penetration of the rays without the invocation of energies at that time deemed prohibitive; (2) the fact that charged particles were believed to possess a definite "range", so that the variation of intensity with altitude had to be accounted for by what appeared to be rather artificial assumptions as to the energy distribution in the incoming rays designed to fit the intensity-altitude curves obtained ; (3) the fact that charged particles, on account of their deflection by the earth's magnetic field, might naturally be supposed, in line with calculations on the subject, to have greater difficulty in reaching the earth's atmosphere for low magnetic latitudes than for high ones.† These considerations led to the expectation of a variation of cosmic-ray intensity with latitude, whereas, at that time, no such variation was known. In 1927, and in subsequent papers in later years, J. Clay (1927–32) published results showing an actual variation with latitude, with an equatorial dip in intensity amounting to about 11 per cent, as the result of experiments made on three different voyages between Java and Holland. On the other hand, Millikan and Cameron (1928), Bothe and Kolhörster (1930), Corlin (1930 a, 30 b) and Grant (1931) failed to confirm these findings in other expeditions in other parts of the world. The matter remained in some state of uncertainty until a world-wide survey by A. H. Compton (1932 a–33 c) and his co-workers (figure 2), supplemented later by numerous observations by other investigators, established the reality of the effect beyond question. The existence of the latitude effect now called as insistently for the recognition of charged particles in the primary radiation as its supposed absence had formerly necessitated a denial of the contribution by such particles.

* A valuable feature of the Geiger counter is its adaptability to use in a form in which two or three such counters are employed in tandem, and associated with circuits in such a manner that only when an ionizing ray passes through all of the counters so employed will the circuits permit a pulse to be recorded. Circuits for the foregoing purpose were first devised by Bothe and Kolhörster (1929), Bothe (1930), Rossi (1930 a), and by Tuve (1930), and since that time all sorts of improvements have been incorporated by numerous investigators. Geiger counters used in the foregoing manner can constitute a "cosmic-ray telescope" for recording intensities in definite directions.

† To these objections a fourth might have been added, to the effect that calculations show that a vertically directed electron at the magnetic equator would have to possess an energy of $1{\cdot}5 \times 10^{10}$ ev., in order to reach the earth's atmosphere in spite of the influence of the earth's magnetic field in tending to return it to space. Thus the only electron rays which could reach the atmosphere in the vertical direction at the magnetic equator would be those possessing more than $1{\cdot}5 \times 10^{10}$ ev., and according to the theories of the day, such rays could penetrate the whole atmosphere, so that no variation of vertical intensity with altitude was to be expected at the equator. Similar considerations follow for other directions and for other kinds of rays.

The foundations pertaining to the effect of the earth's magnetic field upon charged particles approaching the earth had been laid down by Stoermer (1904, 13, 31–34) as early as 1904, but the complexity of the considerations left the picture in a form not readily adaptable to the cosmic-ray considerations. A paper by Lemaître and Vallarta (1933) served to clarify the situation considerably.* By an ingenious application of a dynamical theorem due to Liouville, they showed, as one of their conclusions, that, in spite of the violent bending of the paths of charged particles by the earth's magnetic field, in the case of a group of rays of fixed energy approaching the earth from infinity in all directions with equal intensity, the rays behaved in such a manner that for any assigned point on the earth's surface either no rays reached that point from an assigned direction or they reached it with the same intensity as they would have done had there been no magnetic field.

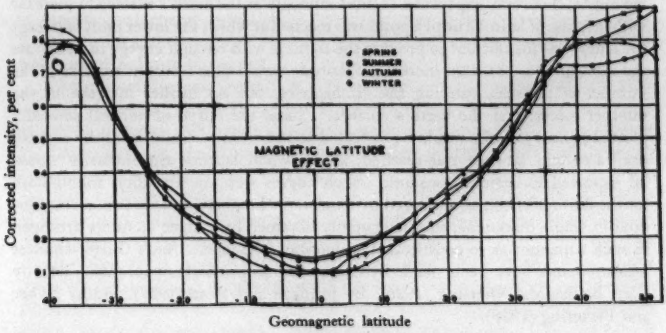


Figure 2.

(Reproduced, by kind permission from the 'Physical Review'.)

The first observations on the latitude effect were made with ionization chambers which integrate the effect from all directions. The results thus yielded do not present the matter in its simplest form, since the effect of variation with latitude is dependent upon the directions in which the rays approach the earth. The general picture is, therefore, most easily discussed in terms of rays of defined direction, let us say vertical rays. Starting then at the magnetic equator, there is a certain minimum energy which the rays must possess to reach the atmosphere. All rays below this energy are excluded, but all rays above this energy not only enter but would have no difficulty in passing right through the atmosphere to the earth's surface if they lost energy only by ionization.

* The theorem of Lemaître and Vallarta was vitiated by an oversight as to the proper interpretation of electromagnetic momentum in the Hamiltonian equations. The analytical effect of this oversight was really the equivalent of assuming absence of any magnetic field. However, a modified demonstration, based on the development by Swann (1933 a) of a theorem closely analogous to Liouville's theorem, removed the defect without destroying the ultimate conclusions.

The problem of what happens to the rays which do enter is one which can only be properly discussed in terms of our present knowledge of all of the atomic processes concerned. In spite of this, however, there is some advantage in considering the story as it had a right to be presented when the only processes of loss of energy were those associated with ionization. The subsequent modifications of the picture will develop naturally from this basis as a starting point. The energy necessary for an electron, for example, to reach the earth vertically at the magnetic equator is $1{\cdot}5 \times 10^{10}$ ev.* On the other hand, the loss of energy by ionization in passing through the atmosphere would be only $1{\cdot}4 \times 10^{9}$ ev. There was thus reason to believe that electrons, and, indeed, any other charged particles, which succeeded in entering at the equator would pass right through the atmosphere. As we progress to higher magnetic latitudes, the lower limit for entry is less, so that more rays might be expected to reach the earth's surface. So long as, in the incoming radiation, there were rays of sufficiently low energy, we might, therefore, expect the vertical intensity at the earth's surface to increase with increase of latitude until a point was reached at which the lower limit of energy for entry was insufficient to provide the particle with enough energy to penetrate the atmosphere. Further increase of latitude would thus result in increase in the number of particles entering the atmosphere, but no further increase in the number received at the earth's surface. Thus the curve of vertical intensity plotted against latitude would be expected to rise from a minimum at the equator until a certain latitude was reached, above which latitude the intensity would be expected to remain constant. Such curves were, in actuality, found both for the latitude effect, as measured by ionization chambers, and for that corresponding to single directions, as subsequently obtained by Geiger counters arranged in such a manner as to constitute a cosmic-ray telescope. Such Geiger-counter measurements have been made by Auger and Leprince-Ringuet (1933, 34), by Clay, Bruins and Wiersma (1936), by Johnson and Read (1937), and by Neher and Pickering (1938).

At the time when these matters seemed to have become clear, the natural charged particles to think of as representative of the cosmic rays were electrons. However, there remained the question as to whether the whole of the radiation was attributable to charged particles. The subsequent developments, as outlined below, served increasingly to confirm the view that charged particles were alone responsible for the primary radiation, but difficulties associated with the assumption that they were electrons, and, indeed, difficulties concerned with the over-all simplicity of the picture, soon made themselves evident. We shall proceed to discuss these matters in sequence.

* The momentum (relativistic) is a quantity more natural to the Hamiltonian equations than is the energy, but discussions are frequently made in terms of the energy. The momentum associated with a charged particle in a magnetic field of vector potential $\mathbf{U}$ is $[m_0 c\beta/(1-\beta^2)^{\frac{1}{2}} - e\mathbf{U}/c]$ where m_0 is the rest mass, β the ratio of the velocity to the light velocity, the units being Heavisidean. At infinity, where $\mathbf{U}$ is zero, the momentum reverts to the ordinary form $m_0 c\beta/(1-\beta^2)^{\frac{1}{2}}$ for a particle in a force-free field, and it is this quantity, in conjunction, of course, with the coordinates, which—when given at infinity—defines, through the Hamiltonian equations, the orbit of the particle, and so the possibility of its reaching the earth. The kinetic energy is $m_0 c^2[1/(1-\beta^2)^{\frac{1}{2}} - 1]$, so that, when β is approximately equal to unity, the energy is proportional to the momentum at infinity.

(iii) *East-west effect and its bearing upon the nature of the charged particle*

The significant elements of the Lemaître-Vallarta theory may be illustrated in the following way* : Suppose that we confine our attention to charged particles of any one energy and, let us say, of positive sign. Then, in general, if we should take our stand at any particular latitude on the surface of the earth, there would be, for any assigned charged-particle energy, a certain cone such that all rays whose directions fell within the angle of that cone could reach the earth with normal intensity. This cone would face in a more or less westerly direction. Outside of the cone, none of the rays could enter. As we moved towards the pole the cone would widen up in angle, so that the direction from which the rays could enter would comprise more and more of the sky until, at the pole itself, the cone would cover the whole of it. Going back to our starting point, however, and moving towards the magnetic equator, the cone would close up, and if the energy were sufficiently low, there is some latitude sufficiently low and such that for that and all lower latitudes the cone would remain completely closed and no rays of the assigned energy could enter. On the other hand, if the energy were sufficiently high, the cone would remain open even at the equator, and to an extent which increased with the energy.

The story of the cone contains the story of the total cosmic-ray intensity with latitude when combined with a law defining the distribution of energy among the incoming particles and with a knowledge of the energy necessary to penetrate the atmosphere, and it also predicts, for particles of one sign, a difference between the intensity from the east and from the west for any given latitude.† If all the rays had one single energy and all carried the same sign of charge, the story of the difference between the intensity from the east and from the west would be very simple. If the rays were positive, and if we were at a place where the cone was not completely open or completely closed, we should observe no intensity from the east and the full intensity from the west. In practice this story is complicated by the existence of an energy spectrum for the incoming rays, the cones being different for each energy. However, the matter is one which lends itself readily to computation of the variation of intensity from east to west in terms of the assigned energy distribution if the rays are all of one sign. It is obvious that if charged particles of both signs occur, the situation will be still further complicated. If the particles of opposite signs existed in equal numbers and were alike in characteristics, there would be no variation in intensity from east to west at all. Following the lines of Johnson's (1938 a) treatment, the relevant matters‡

* The matter is a little more complicated in detail. Thus, between the above two cones in the simplified description there lies a "penumbra" which is crossed by alternating bands of allowed and forbidden directions. Then there is a shadow cone which lies close to the horizon adjacent to the nearer pole and within which lie only orbits—truly of a hypothetical kind—but which would exist if the particles could pass right through the earth, and would then consist of particles which had passed one or more times through the earth before their arrival at the point in question. These and other matters are summarized in a report by Johnson (1938 a).

† The east-west effect was sought, but with a negative result, by Rossi (1930 b–31 b). A positive result was obtained by Johnson and Street (1933) and was later confirmed by more refined measurements by Johnson (1933) and by Alvarez and Compton (1933). Since that time it has been confirmed on many occasions at sea-level and other altitudes.

‡ There is also a north-south effect, and, indeed, certain other effects of analogous kind (see Johnson's report).

in this connection may be summarized as follows : It readily results from the theory, and, indeed, is practically obvious, that the latitude effect at a given altitude depends upon the sum of the intensities contributed at that altitude * by the sum of the incoming positive and negative intensities, while the east-west effect depends upon the difference of those intensities, in both cases over the energy ranges concerned, so that a combination of the data for the two effects enables the relative proportions of each intensity involved to be calculated. By applying these considerations to the data, Johnson concluded that, at any rate over the energy ranges concerned, the incoming radiation is preponderatingly positive.

(iv) *Storm effects and allied phenomena*

Apart from what may be called the normal geomagnetic effects treated above, we have evidence of large cosmic-ray intensity variations produced by magnetic storms and also during the display of the *Aurora Borealis*. Forbush (1937, 38) and Hess (1937) and his co-workers† have found fluctuations of the cosmic radiation occurring simultaneously all over the globe. Such fluctuations constitute another proof of the extra-terrestrial origin of the cosmic rays. These fluctuations can be correlated to some extent with changes of the horizontal intensity of the earth's magnetic field. At the onset of a magnetic storm, a decrease of 5 and even 7 per cent of the cosmic-ray intensity occurs when the horizontal magnetic field increases by less than 1 per cent. This correlation has been discussed by Johnson (1938 a), who raised quantitative objections to the magnitude of the effect. However, Hess considers the reality of the magnetic storm effect to be beyond question.

§ 4. EXPERIMENTS AND CONCEPTS OF THE PRESENT EPOCH

(i) *Difficulties formerly resulting from the charged particle hypothesis*

At the stage and epoch represented by the discovery and interpretation of the geomagnetic effects, the simplest picture sufficient to harmonize phenomena at sea level would be that of an incoming charged-particle radiation of positive sign passing through the atmosphere without scattering or secondary emission other than that represented by ionization, and detectable either in Geiger counters or in ionization chambers by virtue of the ionization produced. However, it soon became necessary to amend this simple picture, and it must be our object to trace the history of such amendments, keeping an eye on the extent to which they may affect those very interpretations of geomagnetic phenomena themselves which led initially to the hypothesis of charged particles of positive sign.

The first difficulty which presents itself is that of providing for a variation of cosmic-ray intensity with altitude in equatorial regions. Confining ourselves for simplicity to the vertical direction, the energy necessary to penetrate the atmosphere is approximately $1{\cdot}4 \times 10^9$ ev. However, at the equator the energy for vertical entry through the earth's magnetic field is approximately $1{\cdot}5 \times 10^{10}$ ev. for an electron, $1{\cdot}4 \times 10^{10}$ ev. for a proton, and $1{\cdot}1 \times 10^{10}$ ev. for an alpha particle, so that, for any charged particle within a reasonable realm of consideration, the energy for vertical entry at the equator would be sufficient to carry the particle

* The formal distinction here implied in speaking of the signs only in relation to the *incoming* radiation is of importance, as will appear later.

† Hess, Demmelmair and Steinmaurer (1938) ; Hess, Steinmaurer and Demmelmair (1938).

right through the atmosphere, and there should be no variation of intensity with altitude at the equator if intensity is measured in terms of numbers of primary rays passing per second through a telescopic apparatus, for example. As the latitude is increased, the condition of no variation of intensity with altitude should persist until a latitude is reached at which the lowest energy for entry is only just sufficient for atmospheric penetration. Considering now any high latitude, there would be, for that latitude, a certain altitude such that all vertical rays entering the atmosphere could penetrate to that altitude, and there would be no variation of vertical intensity with altitude above this critical altitude. As we descended below it, however, we would enter the region to which some of the rays which entered the atmosphere could not penetrate, and the further we descended below the critical altitude the less rays would be received, so that there would be a variation of intensity with altitude below the critical altitude. Obviously the critical altitude would increase with latitude.

Now the situation presented by the above picture, a situation in which there is no variation of vertical intensity with altitude in equatorial regions and only a limited range of altitude for which there is a variation in vertical intensity in the higher latitudes, is one which is absolutely contrary to our experience. It was for this reason that Swann (1934 a, 36 a, 38 a) proposed a hypothesis, new at the time, and to the effect that the primary rays emitted long-range secondaries, the probability of emission of such a secondary being proportional to the energy of the primary. It was not assumed that a secondary emission occurred necessarily in a single act, but possibly through intermediary processes. It appeared that all particles of energies comparable with those concerned in cosmic-ray phenomena, and so the long-range secondaries above referred to, may be expected to perpetuate the path of the primary with but little deviation. The long-range secondaries were, of course, contemplated as electrons, the range being of the order of a few metres in air at atmospheric pressure. They and the primaries were regarded as responsible for operating Geiger counters and producing ionization in ionization chambers, so that this view gave rise to a situation in which that which was measured as cosmic-ray intensity increased with altitude even in equatorial regions, because of the increase with altitude of the energy of the primaries. Since the primaries formed a kind of backbone, geomagnetic effects fundamentally originating in them might be expected to be transferred to the secondaries.

About the time when the above-mentioned hypothesis of secondary production was made, Bethe and Heitler (1934) developed on the basis of modern quantum considerations, but without relation to cosmic rays, a process very like it, but one which was applicable in significant degree only to electrons. This process, born of the quantum theory, became invoked later in that form for cosmic-ray theories and associated matters by several investigators,* but it appeared that in relation to the cosmic radiation itself, its range of applicability was to be found mainly in the higher regions of the atmosphere for reasons which will be evident from its nature. The fundamental ideas characteristic of this new atomic process are as follows : In the first place, the process becomes rapidly applicable for air

* In particular, Oppenheimer (1935), Bhabha and Heitler (1936), Nordheim (1936), Carlson and Oppenheimer (1937), Furry (1937). The fundamentals of the quantum theory processes involved are given in *Quantum Theory of Radiation* by Heitler (1936).

as energies become greater than about $1{\cdot}5 \times 10^8$ ev. Below such energies, the probabilities of occurrence of the essential elements of the process become rapidly evanescent. In the case of a high-energy electron, however, there is a reasonably high probability of the creation of a high-energy photon as the electron passes through a substance, the energy of the electron being, of course, decreased to provide for the energy of the photon. Then the photon, if of high energy, possesses high probability of making a sufficiently close nuclear collision to result in a process formerly completely unknown, namely, the complete disappearance of the photon and the appearance in its place of two electrons, one positive and one negative, with a sharing of the energy between the two determined by appropriate statistical considerations. The positive and negative electrons repeat the process of ejection of photons from atoms, which photons again repeat the process of materialization into other positive and negative electrons.

Now if the processes aforesaid were the only ones which occurred, we should expect that an electron or a photon, starting at the top of the atmosphere or coming from outside it, would result in an ever-increasing line of progeny, so that the charged particle and photon intensity might be expected to increase as we descended into the atmosphere. However, the charged particles are continually losing energy by ionization, and the positive electrons disappear by recombining with electrons in atoms to form photons. Moreover, some of the newly-born charged particles have energies below that appropriate for perpetuating the process, or fall to such energies through ionization. Thus we have a birth-rate of positive and negative electrons and a death-rate, the latter being enhanced by the ionization phenomenon. Starting with a high-energy electron or photon, the birth-rate at first exceeds the death-rate because the energies of the first-born progeny are on the average high and such as to continue the process. However, with successive generations the general average of the energy level falls on account of the sharing of the energy among the many progeny and on account of ionization, so that the death-rate increases and eventually exceeds the birth-rate. Thus the initial high-energy electron or photon is responsible for a contribution to the cosmic-ray intensity, which first increases as we descend into the atmosphere, finally reaches a maximum, and then decreases, the distance of the maximum from the top of the homogeneous atmosphere being greater the greater the energy of the initial particle, electron or photon. The existence of such a maximum and its actual realization for the case of individual rays counted by Geiger-counter telescopes was established about the same time by sounding balloon observations made by Regener and Pfotzer (1935), figure 3, and by the observations made on the second National Geographic U.S. Army Air Corps Stratosphere Flight by Swann, Locher and Danforth (1936 a, 36 b, 37), figure 4. The maximum was also shown later by ionization-chamber data obtained by Bowen, Millikan and Neher (1938), figure 5, and it has been confirmed on many subsequent occasions, as, for example, in the experiments of Johnson (1938 b), figure 6, who used sounding balloons equipped with Geiger-counter telescopes arranged to record their data through radio transmission to a ground station.*

* Accounts of the technique in such measurements have been given by Johnson (1937) and by Clarke and Korff (1941 a, 41 b).

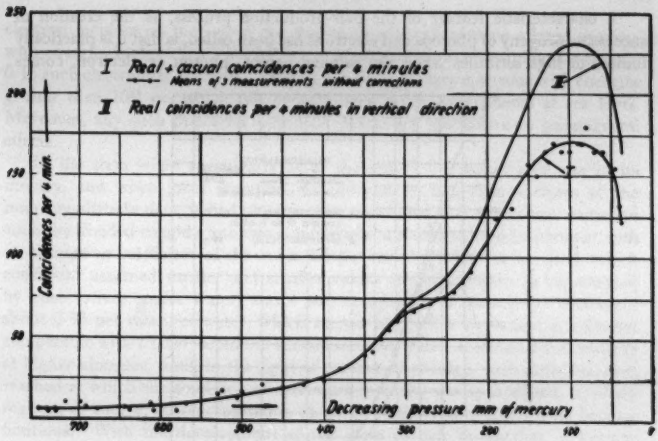


Figure 3.
(*Reproduced, by kind permission, from* '*Nature*'.)

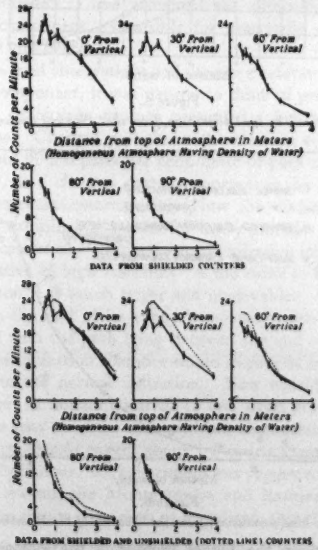

0° From Vertical
30° From Vertical
60° From Vertical
Distance from top of Atmosphere in Meters
(Homogeneous Atmosphere Having Density of Water)
Number of Counts per Minute
80° From Vertical
90° From Vertical
DATA FROM SHIELDED COUNTERS
0° From Vertical
30° From Vertical
60° From Vertical
Distance from top of Atmosphere in Meters
(Homogeneous Atmosphere Having Density of Water)
Number of Counts per Minute
80° From Vertical
90° From Vertical
DATA FROM SHIELDED AND UNSHIELDED (DOTTED LINE) COUNTERS

Figure 4.
(*Reproduced by kind permission of the Council of the National Geographic Society.*)

A characteristic feature of the pair-production process, as the creation of successive progeny of photons and electrons has been called, is that it is practically limited to high altitudes when the original parent, photon or electron, comes

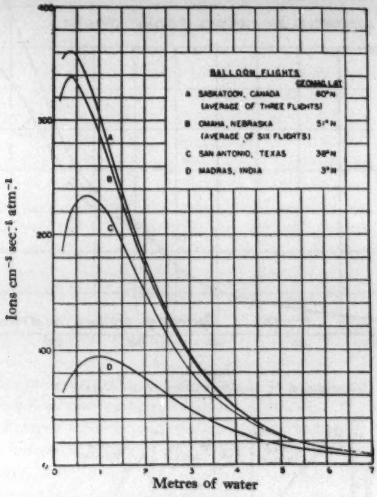


Figure 5.

(*Reproduced, by kind permission, from the 'Physical Review'.*)

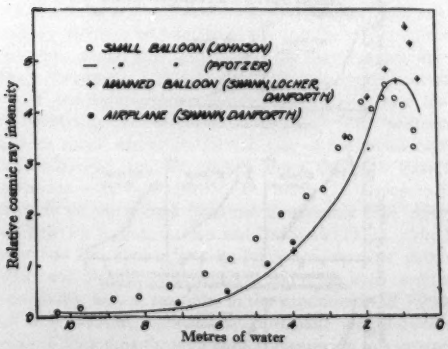


Figure 6.

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

from outside or starts in the higher regions of the atmosphere. Thus, the calculations of Carlson and Oppenheimer (1937) show that, while a 2×10^{11} ev

electron incident on the top of the atmosphere would produce a large progeny with energies higher than $5{\cdot}5 \times 10^8$ ev., it would, on the average, give rise to only 0·15 such electrons at sea level. Indeed, it is only primary electrons with energies greater than 10^{13} ev. which can provide even a few descendents at sea level. Moreover, any such primaries would be practically insensitive to geomagnetic effects.

In the days when speculation as to the nature of cosmic rays was in its infancy, and when great attention was devoted to empirical analyses of the intensity-altitude data, it had already been concluded that the cosmic radiation could be divided roughly into two components, a so-called hard component with an absorption coefficient of the order 0·1 per metre of water at sea level, which coefficient assumed smaller and smaller values at lower depths as represented by observations under water, and a soft component of absorption coefficient about 0·55 per metre of water, which started to become important at altitudes comparable with 15,000 feet in the atmosphere and which dominated the situation at higher altitudes, until, in the light of present knowledge, such altitudes were reached at which the condition of maximum intensity was approached, in which regions, of course, all representation in terms of absorption coefficients became fictitious. With the development of the ideas of pair production, it became natural to associate what had heretofore been called the soft component with this process and to see that component as originating from electrons entering our atmosphere from outside. There remained the difficulty of accounting for the so-called hard component responsible for practically everything that was observed at sea level and below.

Recalling that sea-level observations and those at moderate altitudes demanded a positive primary component, it was natural to think of protons as responsible for this component. However, protons presented a fundamental difficulty for the following reasons. Charged particles with velocities comparable with that of light produce about 30 ion pairs per centimetre of path in air at atmospheric pressure. However, as a charged particle approaches the end of its range, its ionization activity increases enormously. Now the distance over which this increased activity is exhibited is greater for heavy particles than for light ones, so that while, in the case of an electron, the total ionization in excess of the normal ionization representative of high velocities * is too small to be readily observed, in the case of a proton it is much larger and observable. In other words, if a proton should reach the end of its range in an ionization chamber, the total ionization associated with the path lying between the end of the range and the point of entry into the ionization chamber would in general be much greater than that corresponding to the normal ionization. Now considering an ionization chamber 45·5 sq. cm. in horizontal area, 15 cm. deep, filled to a pressure of 10 atmospheres, it is easy to show that, with an absorption coefficient of 0·08 per metre of water, we should expect about 1·2 spurts greater than 54,000 ions to occur in this vessel per hour. An experiment was suggested by Swann (1936 b), and carried out by Swann, the Montgomerys and Ramsey (1936), to test for such an effect, and the results were such as to indicate that the number of spurts

* For extremely high energies the ionization per centimetre of path increases slowly with the energy. We may designate the normal ionization as that corresponding to the minimum of the ionization energy curve.

of ionization of the kind contemplated was far too small to be consistent with the assumption that the rays observable at sea level are primarily protons.

It was about this time that Street and Stevenson (1937a) found, in cloud-chamber observations, particles whose deviation in a magnetic field could only be reconciled with their ionization per cm. of path, as measured by the droplets in the cloud chamber, by the assumption that they were particles with a mass of the order 150 times that of the electron.* Both signs of charge are represented in these particles, and they came to be known by various names from which we shall choose the designation, mesotron. It was soon found that these new particles possessed a characteristic unknown in any other kind of atomic particle, the characteristic of possessing a finite life. This conclusion emerged from experiments which took various forms, but which can all be thought of in terms of an ideal experiment in which it becomes shown that the absorption of mesotrons in a compressed column of air is less than it would be if the column were uncompressed and were, therefore, much longer. The number of processes associated with ionization, and, indeed, any other atomic phenomenon, would be the same for a ray passing through either column, and any difference in absorption, therefore, is most naturally attributable to a pure time effect, such as would occur if the mesotron had a finite life. Experiments founded upon the equivalent of this idea have shown that a mesotron has an average life comparable with 2×10^{-6} sec. when it is moving with a very slow velocity—theoretically, when it is at rest.†

Wave-mechanical criteria show that pair-formation phenomena become less important as the rest mass of a particle is increased, and that while they are very important for electrons, they would be of negligible account for mesotrons and, of course, of still less account for protons. Thus the mesotron loses energy primarily by ionization, and on this account possesses a much larger potentiality for penetration than would an electron of the same energy. A mesotron with an energy of $1{\cdot}4 \times 10^{11}$ ev. could penetrate through an equivalent of 1000 m. of water.

At this stage, therefore, it might seem that a ready basis was provided for the main phenomena in assuming that what had been regarded as the soft component was attributable mainly to primary electrons which, through pair formation, gave rise to the greater part of what was observed in the higher regions of the

* Previously, Anderson and Neddermeyer (1936) had shown that for energies up to 3×10^8 to 4×10^8 ev. the cosmic-ray shower particles have energy losses in lead plates corresponding to those predicted by quantum mechanics for electrons. Later studies of range by Street and Stevenson (1937 b) and of energy loss by Neddermeyer and Anderson (1937) indicated that the singly occurring cosmic-ray particles, even in the range below 4×10^8 ev., are more penetrating than the shower particles of corresponding magnetic deflection. Further discussion of this matter is contained in a paper by Neddermeyer and Anderson (1939). A summary and discussion of information pertaining to the mesotron, with many references, has been given by Peierls (1939).

† If we should suppose that the lifetime of a mesotron (of velocity βc), as measured by one who accompanies it in its motion, is τ_0, then it follows immediately from the relativity transformation that the lifetime as measured by a stationary observer is τ, where

$$\tau = \tau_0/(1-\beta^2)^{\frac{1}{2}}.$$

Thus, since the kinetic energy of the particle T is given by

$$T = m_0 c^2[1/(1-\beta^2)^{\frac{1}{2}} - 1],$$

where m_0 is the rest mass and c the velocity of light, we see that the life of the mesotron increases with its kinetic energy, and for high values of that energy is proportional to it.

atmosphere (above 15,000 feet, for example), and that the so-called hard component was attributable to mesotrons. However, such an assumption immediately invited two fundamental difficulties. The life of a slowly moving mesotron is only of the order of 2×10^{-6} sec. A mesotron with high energy has a lifetime which is greater than this in the ratio 1 to $(1-\beta^2)^{\frac{1}{2}}$. The ratio in question is the ratio of the actual mass to the rest mass, and for high energies is the ratio of the actual kinetic energy to m_0c^2, where m_0 is the rest mass. For mesotrons of 10^{10} ev. energy, this ratio would amount to only about 100, so that the lifetime would be of the order of 2×10^{-4} sec., which would permit a particle to travel on the average only about 60 km. before terminating its existence. Even if the mesotrons started from a body as near as the sun, therefore, they would have no possibility of reaching the earth. The only conclusion is that they must be born in the atmosphere itself, and, presumably, in the higher regions of the atmosphere.*

Another difficulty confronting the assumption of the external origin of the mesotrons lies in the fact, as Johnson's (1938 a) results, and those of others, have shown, that practically all the incoming particles which are responsible for effects at sea level are of positive sign. On the other hand, cloud-chamber observations have shown that mesotrons are distributed practically equally in sign. This difficulty, as well as the former difficulty, is met if we assume that the primary radiation is a positively-charged particle which, by processes as yet not thoroughly understood, gives rise on entering the atmosphere to mesotrons. The high-energy mesotrons which are capable of reaching the earth, in virtue of that high energy, perpetuate the paths of the positive primaries in spite of the fact that they themselves may be equally distributed in sign.

Once again, consistency of interpretation of the main results might seem to have been secured by the assumption that the soft component is attributable to electrons, as aforesaid, and the hard component to mesotrons born of some positive primary in the higher regions of the atmosphere. However, another element enters to disturb the simplicity of this solution.

In the National Geographic U.S. Army Air Corps stratosphere flights made by *Explorer I* and *Explorer II*, and in the Jean Piccard flight, Swann and Locher (1935) and Swann, Locher and Danforth (1936 a–37) made observations of the variation of cosmic-ray intensity with altitude for various zenith angles. It was found that at the higher altitudes the diminution of intensity with increase of zenith angle from the vertical was much less rapid than would be expected if the intensity in any direction were determined simply by the line integral of the atmospheric density from the point of observation upwards to infinity along the direction concerned.† Figure 7 shows the number of impulses recorded per minute by the telescopes plotted against the zenith angle for these flights.

* A further argument against primary mesotrons lies in the fact that when we measure the energies of mesotrons at sea level, and then, through calculation of their energy-loss by ionization, deduce the energies which they must have at the top of the atmosphere, these values are in many cases too small to permit them to have entered through the earth's magnetic field.

† The assumption that the intensity is determined entirely by the aforesaid line integral is usually stated by asserting that the intensity is a function only of $h \sec \theta$, where θ is the zenith angle. This law is implied in all such analyses as those of Gross (1933).

It was found that in the flight of *Explorer I*, at an atmospheric depth equivalent to 2 m. of water, the impulse counts for the horizontal direction formed 25 per cent of the corresponding value for the vertical. In the Piccard flight, at a water equivalent depth of 1 m., the corresponding ratio was 50 per cent, and in the flight of *Explorer II*, at a water equivalent depth of 0·5 m., it was 80 per cent. Even allowing for the fact that a horizontal telescope counts from both ends, whereas a vertical telescope—except under extraordinary conditions—counts only from the top end, we have far less variation of intensity with zenith angle than can be accounted for by a radiation coming from outside and being absorbed to an extent depending upon the mass, per unit area, of atmosphere traversed.*

At first sight one might think that at a water equivalent depth of half a metre the line integral of the density to infinity in a direction tangential to the horizontal at the point of observation might be so small as to account for the large horizontal

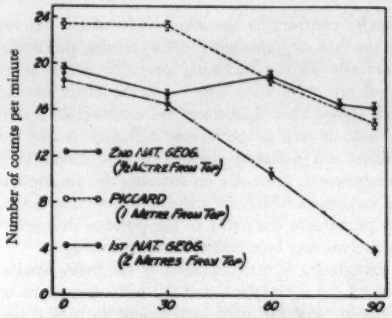


Figure 7.

(*Reproduced by kind permission of the Council of the National Geographic Society.*)

intensity. That such is not the case is borne out by the following considerations. Suppose we consider a direction defined by a zenith angle† θ given by $\cos\theta = 0{\cdot}1$, the point of observation being at a water equivalent depth of 0·5 m. On the basis of the assumption of an absorption depending upon the line integral of the density, we should expect, for a direction given by $\cos\theta = 0{\cdot}1$, an intensity equal to the vertical intensity at a depth of ten times 0·5 m. of water. The vertical intensity curves show that the vertical intensity at 0·5 m. depth is about six times that at 5 m. depth. However, the intensity as measured for the direction given by $\cos\theta = 0{\cdot}1$ differs from the measured vertical intensity at the same altitude (0·5 m. of water) by less than 20 per cent. Hence, the zenith-angle results cannot be accounted for on the basis of an absorption depending only upon the line integral of the density. A further support of these conclusions is provided by the fact that even at the highest altitudes obtained by *Explorer II* the intensity from the east differed from that from the west by less than 1 per cent, and both

* The assumption of a mean life for the particles, as in the case of mesotrons, would sharpen the variation of intensity with zenith angle.

† The choice of a zenith angle not quite 90° avoids the irrelevant complications introduced by the curvature of the earth.

intensities differed from the average for all azimuths by less than 1 per cent. Had the electrons come from outside the atmosphere, one would have expected an east-west magnetic effect. Even stronger evidence of the same kind was supplied by the experiments of Johnson and Barry (1939), made at high altitudes and low latitudes, which gave no evidence of any variation of intensity with azimuth in spite of the fact that large variations of this kind might have been expected had the electrons come from outside.

On the basis of the foregoing considerations, Swann (1939 a–41) concluded that a large part of the electron intensity found at very high altitudes must be attributable to electrons having their origin in the atmosphere itself, and it is natural to suppose that all may have their origin in this manner.*

It is possible to account for the uniformity of intensity in all directions in the stratosphere on the basis of the following assumptions :—

Let us suppose that a charged primary enters the atmosphere and then breaks up into mesotrons in such a manner that the resulting mesotrons continue to travel along with the velocity which the primary possessed before the catastrophe. In other words, the assumption is that all that happens is that the heavy particle splits up so that its mass is divided between the mesotrons, and so, consequently, that the velocity of the mesotron is equal to that of the primary. Now if the primary particle is sufficiently heavy, the lower limit of its velocity for entry through the magnetic field will be low, and the velocity of the mesotrons born of such a primary will be low. A sufficiently slow mesotron has a mean life of only 2×10^{-6} sec., and it could not travel more than 600 m. on the average without disintegrating, since it must travel with a velocity less than that of light. It is customary to assume that, at the end of its life, a mesotron disintegrates into an electron and a neutrino, and these may be regarded as being emitted, on the average, equally in all directions in relation to axes moving with the mesotron. For a low-velocity mesotron there is equality of emission in all directions in terms of stationary axes. Thus, in so far as the intensity in the stratosphere is contributed to by electrons born from mesotrons which are the offspring of low-velocity primaries, we may expect equality of emission in all directions, even in the upward direction, as the contribution from such electrons.

As we descend from the higher regions of the stratosphere, we may expect to find the homogeneity of the radiation in all directions becoming diminished because the mesotrons which reach to these levels are those which have lived long enough to do so. In other words, since mesotron life increases with kinetic energy, these mesotrons must be those of velocity comparable with that of light. In the system of axes moving with one of these mesotrons, the disintegration of the mesotron may occur with equal probability in any direction, but when we transfer this story to fixed axes we find a favouring of the forward direction for the electron emission. Thus, as we descend through the stratosphere into the lower regions of the atmosphere, we may expect the symmetry of emission to diminish, until finally we approximate to the intensity-zenith angle curves found at low altitudes. In other words, the conclusions are in harmony with our experimental findings.

* Similar conclusions were reached later by Schein, Jesse, and Wollan (1941), as a result of other considerations.

(ii) *Determination of the mass of the primary or primaries*

The foregoing considerations present a general picture, but Swann (1943) has shown that it is possible to go further in such a manner as to utilize the experimental data to the end of deciding the nature of the primary particle itself. It turns out, in fact, that two types of primary particles are required, the lighter particle, a proton which, through its progeny, is responsible for effects below the stratosphere, and a heavier particle which, through its progeny, is dominant in the high regions of the stratosphere. The bases of these conclusions are as follows :—

As a first approximation, we assume, as seems quite customary in this field, that there is an effective altitude of production—a kind of centre of gravity of production. We assume, moreover, a spectral distribution law according to which the number of primaries *which, through their progeny, are responsible for effects near sea level*, and whose energies lie between E and $E+dE$, is proportional to $1/E^3$, with, of course, a lower limit of applicability below that for entry through the magnetic field.* In virtue of the assumption made as to the splitting of the primaries into mesotrons, the distribution of kinetic energy among the mesotrons in the layer of production will be exactly the same as that among the primaries, except that corresponding energies in the case of the mesotrons will all be divided by N, where N is the number of mesotrons produced by a primary and, therefore, the ratio of the mass of the primary to that of the mesotron.

Taking into account the known energy-loss by ionization in the passage of mesotrons through the atmosphere, the mean life of the mesotron and its variation with energy, it is then possible to work out what is to be expected as regards the variation of mesotron intensity with altitude and latitude in the vicinity of sea level. It is then possible to adjust the assumptions as regards altitude of production to harmonize with the facts. Then, from a knowledge of the ratio of the vertical intensity at two latitudes, one of which is that of the equator, it is possible to determine, for the equator, the *lower limit of the kinetic energy* for the mesotrons at the altitude of production. This yields *the lower limit of the velocity of the mesotrons, which is also the lower limit of the velocity for the heavy particles.* We can then ask ourselves what mass the heavy particle would have to possess in order that, with this velocity, it could just have penetrated the earth's magnetic field at the equator. The result comes out *equal to the mass of the proton.* The observations used in the above calculations are those at sea level and at altitudes not above four kilometres, so that we can conclude that the mesotrons observed *at these altitudes* have come from primary particles which are *protons.* It must be added that the idea of the primary consisting, at any rate in part, of a heavy particle has been mooted by several investigators, and naturally the proton has been suggested as a likely candidate for the position.† The primary purpose of

* Zanstra (1934) has given a method of deriving from the latitude effect the form of the energy-distribution functions, which method is applicable, within limits and with suitable adjustments, when the primaries themselves do not penetrate the atmosphere but hand on their responsibility to long-range secondaries. An energy distribution involving an inverse power approximately equal to 3 has been reached as a result of this and of information obtained from other cosmic-ray phenomena, such as variation of intensity with altitude, etc.

† The possibilities inherent in heavy particle primaries, and protons in particular, have been considered by Compton and Bethe (1934), Johnson (1934), Compton (1936 a), Johnson (1939), Swann (1939 a–43), Schein, Jesse and Wollan (1941). As a rule, a primary electron component has been assumed in addition.

the calculations just cited was to *determine* the mass of the primary particle which is responsible for effects near sea level and show that it is, in actuality, a proton. This matter has some importance, for there is evidence that the proton is not the only heavy primary particle, as will appear later.

It is of interest to observe that the altitude of production yielded by the foregoing considerations is 16·6 km., which is in good agreement with the altitude (14·2 km.) found experimentally by Schein * (1944) and his collaborators for the maximum of the intensity-altitude curve for mesotrons.

(iii.) *A possible helium component*

We have already referred to the necessity of assuming that the electrons operative in the higher stratosphere originate in the atmosphere rather than at some place external thereto, and it is natural to assume that they result from the death of those mesotrons which, born of the death of the protons, have a sufficiently low energy to have, in consequence, but a short life. There is, however, a difficulty in this assumption ; for, having committed ourselves to the proton as a primary particle, we are committed to a lower limit for the velocity of entry at any latitude and so to a lower limit for the mesotron. Now the extent to which the electrons born of the mesotrons are shot out uniformly in all directions as viewed from stationary axes is dependent upon the velocity of the parent mesotrons. If the velocity is too high, there will be a considerable perpetuation of the mesotron's direction of flight by the electron. Now the lower limit of velocity assignable to the mesotrons on the assumption that they are the offspring of protons with a lower limit of velocity for the entry of such particles results in a condition in which we cannot explain the considerable intensity found in the stratosphere at reasonably large zenith angles. If we assume the existence of a particle heavier than a proton, its energy for entry will carry with it a lower velocity than for the proton and will provide for correspondingly lower-energy mesotrons. If we assume that the incoming particle is a singly-ionized helium atom, it turns out that the lower limit of energy of the mesotrons is such as to provide for electrons whose directions depart appreciably from those of the mesotrons themselves and such that, on carrying through the calculation on the basis of reasonable assumptions, we find that the ratio of the horizontal to the vertical intensity due to the electrons should be 0·22 and 0·13 for altitudes at which the atmospheric pressure is 0·1 and 0·2 atmosphere respectively. The experimental data taken from the flights already referred to yield for the aforesaid ratio 0·21 and 0·13 respectively for the two altitudes in question.† The closeness of the agreement is, of course, to some extent accidental in view of the nature of the computations. Thus the direct mesotron contribution to the intensity is omitted in the calculations. However, according to the conclusions of Schein, Jesse, and Wollan (1940), about 88 per cent of the radiation at these high altitudes is to be attributed to electrons.‡ Then again, the contribution by the electrons

* Given in a paper presented before the American Physical Society at Pittsburgh, April 28, 1944. This paper is given in title only in the *Physical Review* and is not abstracted.

† The details of this matter are given in the publication already cited—Swann (1943).

‡ One must observe that although the death of a mesotron gives rise, in the first instance, only to one electron, that electron becomes multiplied severalfold by pair production.

born indirectly from the protons is neglected, and the inference to be drawn from the agreement above cited is to the effect that the helium-born electrons are predominant at very high altitudes.

One might wonder how far the invocation of helium atoms was necessary and whether, by a modification of the hypothesis as to the production of mesotrons from protons, it would be possible to account for the high intensities at large zenith angles on the basis of protons as the only original parents. One cannot escape the necessity of the assumption of mesotrons of lower mass ratio * than would follow from the hypothesis that all the mesotron offspring of a parent moved with the velocity of that parent if the parent were a proton. The desired conditions of a low mass ratio for some of the proton-born mesotrons could be secured if, in the system of the moving proton, the mesotrons were born with unequal velocities. † Under such a hypothesis, one would only secure a small mass ratio for some of the mesotrons at the expense of a larger mass ratio for others, and it is doubtful if the gain secured in this way would be sufficient.

It is true that in the present state of our knowledge of atomic processes associated with the creation of mesotrons we cannot be very certain of our ground. Thus, several entities and processes, whose existence can hardly be claimed to have been established, stand as candidates for a rôle in the phenomena.‡ Arley and Heitler (1938) have proposed the existence of neutral mesotrons which have an infinite mean life until they encounter matter, when they may give rise to ordinary charged mesotrons with finite mean life. Such "neutrettos", as they have been called, could function in part as primaries, and could enter through the earth's magnetic field without hindrance. They could account for observations such as those of Schein, Jesse and Wollan (1940), in which mesotron production by a non-ionizing radiation was observed. Such considerations as these may conceivably contain an explanation of the low-energy mesotrons necessary to provide for the intensity-zenith angle considerations cited above, and without the invocation of the helium hypothesis. However, in competition with that hypothesis, they have to face the fact that helium, the next known particle to the proton in the ordinary scale of mass, fits the facts so remarkably well, without any artificial adjustments of the primary and rather simple hypothesis so successful in yielding the proton mass for the particle which is the parent of the particles responsible for effects at sea level.

(iv) *Variation of the soft component with altitude*

On the view that electrons in the cosmic radiation arise for the most part from the death of mesotrons, the story of the mesotron disintegration should contain within itself that of the variation of electron intensity with altitude.

Figure 8 shows, according to Schein, Jesse and Wollan (1940), the mesotron intensity curve obtained by these investigators and the total vertical intensity

* That is, ratio of actual mass to rest mass.

† Viewed from the frame of reference of the proton, this would imply that kinetic energy was created. Such kinetic energy would have to come either from nuclear encounters with atoms of the air which, in this frame, are moving rapidly past the proton, or from a disappearance of a certain amount of rest mass in the form of energy.

‡ Accounts of some of these matters are given by Schein and Wilson (1939 a), by Nordheim and Hebb (1939), by Heisenberg (1939), by Peierls (1939), and by Hamilton, Heitler and Peng (1943).

curve obtained by Pfotzer. The curve for electron intensity as a function of altitude is obtained by the aforesaid authors by subtracting these two curves, as is shown in figure 8. It rises much more rapidly with altitude than does the

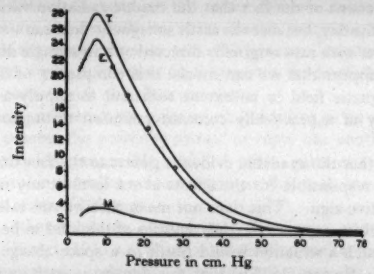


Figure 8. The plot represents, as a function of pressure, T the total vertical intensity as measured by Pfotzer, M the vertical mesotron intensity from present experiments, E the difference between curves T and M, which should give the vertical electron intensity. The open circles, which represent the shower intensity as measured by Regener and Ehmert, are seen to follow very closely curve E.
(*Reproduced, by kind permission, from the 'Physical Review'.*)

mesotron curve. We can understand this situation by observing that, as we ascend, the ionization loss per centimetre of path decreases on account of the decrease of atmospheric pressure. Thus, if electrons are *produced* at all altitudes in proportion to the mesotron intensity, the electron progeny per mesotron will *build up* more at high than at low altitudes.*

§5. THE ORIGIN OF THE COSMIC RADIATION

The difficulty of deciding upon the origin of the cosmic radiation lies not so much in inventing a hypothesis which would be satisfactory for the purpose, as in verifying whether or not the said hypothesis is representative of what occurs in nature. It is certain that the cosmic rays originate in places which we cannot reach to investigate the details of their birth, and, barring such a test, therefore, many hypotheses may stand with equal rights for consideration. The ideal solution will be reached if and when some process in astronomical physics, capable of giving rise to the cosmic radiation, is found to be a *necessary* consequence of other known astronomical phenomena. Under such conditions we should not have to limit ourselves to a statement that the cosmic radiation *could* originate in such and such a manner, but should have available the much stronger conclusion that such and such a process is known to occur and it *must inevitably* produce cosmic rays in the desired quantity. Until this desired consummation has been attained, we can but limit ourselves to discussing the various processes which might be conceived as reasonable of happening in nature and which could produce cosmic rays.

* This matter has been examined in detail by Rossi (1940) and by Rossi and Greisen (1942), with the conclusion that decay electrons are not sufficient to explain the whole of the soft component. They attribute the remainder to electrons which are the descendants of very high-energy primary electrons which form the basis of the so-called extensive showers referred to in Part II of this *Report*.

The sun as a possible origin naturally stands first in consideration, particularly as the related phenomenon of the *Aurora Borealis* appears certainly to be concerned with charged particles from the sun. We cannot immediately rule the sun out on account of the fact that the cosmic radiation varies very little, if at all, throughout the day, because the earth's magnetic field can succeed in bathing the earth at all parts with rays originally directed at it in a single direction. However, it does not appear that we can invoke this complexity of distortion of ray paths by the magnetic field to an extent sufficient to supply a reason for the extreme simplicity of a practically constant radiation throughout the 24-hour cycle.

We have seen that the available evidence points to the fact that the incoming primary radiation responsible for the effects at sea level at any rate is predominantly of the positive sign. This does not mean that we can tolerate a situation in which these positive rays are the only entities of the kind to be found in interstellar space, for such a situation would result in a space-charge density, and it has been shown by Swann (1933 b) that this density is such as would give rise to potential differences of the order 10^{18} volts between points a light year apart. Hence it has become necessary to suppose the positive rays neutralized by a negative space charge of slowly moving particles in interstellar space.

Quite apart from the reasons already given for supposing that electrons do not participate in the primary radiation, it has to be recognized that electrons coming from bodies such as stars, for example, would have to pass through considerable thicknesses of matter on their way out, and apart from the difficulties in emerging thus encountered, we should expect that those which did emerge would be accompanied in approximately equal numbers and energy distribution by photons produced in the journey. The experimental evidence gives little support to the existence of such photons.

Compton and Getting (1935), and Compton (1936 b) have called attention to the fact that if the cosmic rays travel symmetrically in all directions in intergalactic space, then, on account of the motion of the earth in our own galaxy, a greater intensity should be experienced at any instant on that side of the earth which is moving forward than on the other side which is moving backwards, for the same reason that a person flying through a symmetrical rain-storm would receive more drops on the forwardly moving surface than on the backwardly moving one.* The argument of Compton is that if an effect of this kind were found—and the preliminary experiments appeared to indicate such an effect—it would be evidence that the rays were symmetrically directed in inter-galactic space and so came from regions outside the rotating galaxy. On the other hand Alfvén (1938) has maintained that our galaxy must have a general overall magnetic field of the order 0·1 gauss, which would tend to imprison cosmic rays within it and complicate considerably the story of any cosmic rays which entered it. Incidentally, it has been maintained that such a magnetic field would relieve hypo-

* The true picture requires modification from this simple one on account of the fact that, with particles moving at velocities practically equal to that of light, relativity considerations demand that the velocity of the rays, as they appear to the observer, must be the same if he were on the forwardly moving side of the earth as if he were on the backwardly moving side. The picture presented by relativity is to the effect that the observed *densities* on the two sides are different. These and other matters concerned with the problem have been discussed by Swann (1937).

thetical primary electrons of the encumbrance of accompanying photons, as cited above, since the latter would escape from the galaxy while the former would tend to be imprisoned by the magnetic field. Alfvén's theory of the existence of the magnetic field is founded on the rather dubious concept that the heavy particles comprising the neutralizing space charge participate in the rotation of the galaxy, while the cosmic rays do not. Moreover, Compton (1939) has given cogent reasons to support the view that magnetic fields of the order 0·1 gauss would be inconsistent with the observed facts.

In connection with magnetic galaxies, it is further to be realized that such galaxies would possess the power to attract or repel one another, and it can be shown that with fields comparable with 0·1 gauss, galaxies at normal distances apart could give to one another velocities comparable with one-tenth that of light in a period of the order of 10^8 years.

Concerning other methods of acquiring cosmic-ray energies, it has been pointed out by Swann (1933 a, 33 b) that fluctuating stellar magnetic fields can, through electromagnetic induction, give rise to cosmic-ray energies.* It is not the magnitude of the field which is important, but its rate of change, and fields fluctuating at the rate comparable with the rate observed in sunspots can, if spread over a sufficiently large area in a star, give rise to energies of the order 10^{10} ev. without the absolute magnitudes of the fields being anything like those comparable to the observed fields in sunspots. Indeed, fields quite unobservable by spectroscopic means would be sufficient.

There is always open the possibility suggested by terrestrial thunderstorms. When we observe that a terrestrial thunderstorm can give rise to potential differences of the order of 10^9 volts, and when we recall that a thunderstorm is nothing more than a huge electrostatic machine operating under the influence of gravity, it would not be unreasonable to contemplate phenomena of the character of "stellar thunderstorms", but not, of course, utilizing water vapour, which could give rise to energies comparable with much more than 10^9 ev.

If the stars possessed electric fields of an amount comparable with the earth's electric field, and provided that these fields extended out to infinity (which the earth's field does not do) according to the law of inverse squares, then, as Swann (1932, 38 a) has pointed out, galaxies composed of many stars might acquire very large average potentials at infinity, potentials of the order of 10^{10} volts. A field *at* one of the stars itself would be much larger. In the case of one of the examples considered, it would amount to 6×10^{12} volts.

Modifications of electrodynamic laws have been suggested by Swann (1927 34 b, 38 a) to explain the earth's magnetic field and the maintenance of its charge in spite of a continual atmospheric electric current to infinity. The space charge generated by such outward flow of current can be shown to result in very large differences of potential under suitable conditions between points in space.

Gunn (1933) has offered an ingenious suggestion to provide for an emission of heavy particles and light particles from the stars under conditions in which both kinds attain the same average velocity. Then the earth's magnetic field acts as a barrier to the particles of low rest mass, admitting the others. With the

* The principle here invoked is the same as that subsequently utilized in the betatron.

assumption that the heavy particles are positively charged, the positive sign of the incoming primary rays is thus accounted for.

Alfvén (1936) has suggested a mechanism in which two magnetic double stars rotating around each other can, by means somewhat analogous to those prevailing in the cyclotron, give rise to energies comparable with cosmic-ray energies. The theory of the process is complicated by certain subtle considerations which have been discussed by Sabato (1939), who considers the hypothesis untenable.

Finally, we have such considerations as those already referred to, and suggested originally by Millikan, according to which the rays originate in intergalactic space by atom-building processes involving the transformation of rest mass into ordinary energy.* Such theories are intriguing from the point of view of numerical relationships drawn from the mass-energy relation, but nuclear physics has not yet reached the stage at which it can speak with much authority to support the energy transformations involved.

§6. MATTERS PERTAINING TO COSMIC-RAY ENERGY

Since all of the cosmic-ray energy entering the atmosphere is ultimately expended in ionization,† it is possible, by integrating the ionization per centimetre of path from any altitude down to sea level and below to the greatest depth at which cosmic radiation can be observed, to obtain the total cosmic-ray energy passing through a sphere concentric with the earth at the altitude in question. By carrying out this operation for an altitude representative of the top of the homogeneous atmosphere, we can obtain the total cosmic-ray ionization entering the earth's atmosphere. If this is divided by the average energy of the entering rays, the result gives the number of rays per second entering the earth's atmosphere. If we assume a certain type of ray, the minimum energy for entry through the earth's magnetic field is calculable and the average energy of those rays which enter is also calculable if we assume a spectral distribution law. Information concerning the spectral distribution law over the range of energies comprised in the latitude effect may be deduced from that effect by Zanstra's (1934) method. A spectral distribution inversely as the third power of the energy fits the fact satisfactorily, and if this law be assumed to extend to higher energies beyond the range of the latitude effect, it is possible to calculate the average energy of the rays and so determine, as aforesaid, the total number of cosmic rays entering the earth. Such a calculation has been made by Johnson (1938c) under the assumption that the energy for entry is that corresponding to electrons, and that the energy for vertical entry may be taken as a mean. The results obtained in this way are approximately valid for other cosmic-ray charged particles.

From the foregoing calculations it results that $8{\cdot}0 \times 10^{17}$ rays enter the earth's atmosphere per second, and the total power of the cosmic-ray primaries is about one million kilowatts. If the rays were all positive, the current to the earth would be 0·13 ampere.

* A more recent form of this hypothesis has been presented by Millikan, Neher and Pickering (1942).

† Unless, of course, there exists a radiation which does not ionize and does not give rise to an ionizing radiation.

PART II. DETAILS AND DISCUSSION OF SPECIFIC COSMIC-RAY APPARATUS, PROBLEMS AND PHENOMENA

§7. THE PRINCIPAL TECHNIQUES OF COSMIC-RAY RESEARCH

(i) *Principal techniques*

It is not practicable to give a complete list of all of the techniques used in cosmic-ray investigations. However, the principal techniques may be classified under the categories: Ionization chambers, Geiger-counter cosmic-ray telescopes, Geiger-counter shower-measuring apparatus, and cloud chambers. All of this apparatus has, of course, assumed special forms for high-altitude work.

(ii) *Ionization chambers*

Frequently ionization chambers have been pressure-filled. Electrical measuring systems have comprised single-fibre electrometers, bi-filar electrometers, other types of fibre electrometers and quadrant electrometers. Figures 9*, 10 and 11 show three types of apparatus founded on ionization chambers, the

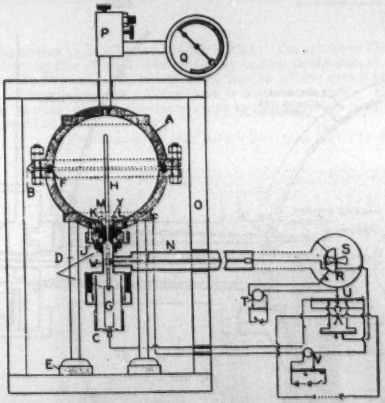
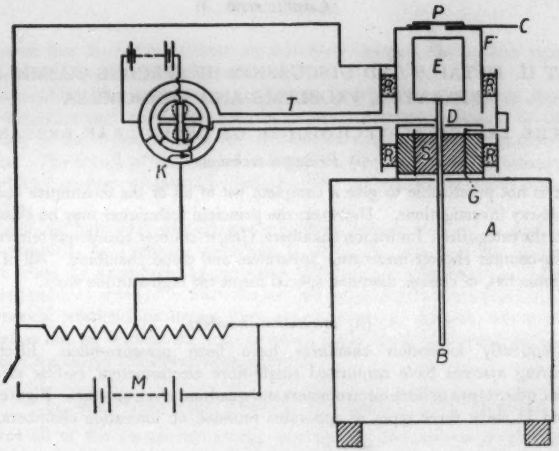


Figure 10. Apparatus by Swann (1930). The pressure-filled iron sphere A surrounds the rod H, insulated from the sphere and protected by a guard-ring. The part G with its surrounding guard-ring protected assembly serves as a compensator, so that fluctuations of the battery applied at the ends of U may be compensated as regards their effect on the quadrant electrometer. The variable potentiometer system V enables observations to be obtained by a null method. The calibrating system T is for subsidiary purposes.

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

apparatus being adapted for ground observations. Figure 12 shows apparatus adapted for high-altitude observations with automatic photographic recording. The details of these pieces of apparatus will be clear from the captions.

* Facing page 28.



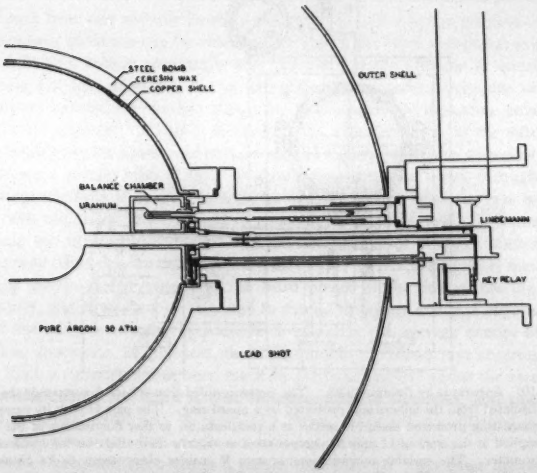


Figure 11. Both pieces of apparatus employ a radioactive source to compensate the ionization current. The upper apparatus, by Swann (1925), employs a 300-litre ionization chamber with central insulated rod protected by a guard-ring and provided with a compensator E to eliminate battery fluctuations as in figure 10. P is the radioactive source, operating through a hole in the variable shutter C.

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

The lower apparatus, by Compton, Wollan and Bennett (1934), employs a 30-litre pressure-filled sphere. A radioactive source in conjunction with a scheme following the general lines of the upper figure provides almost complete compensation.

(*Reproduced, by kind permission, from the 'Review of Scientific Instruments'.*)

In both pieces of apparatus the uncompensated current is observed—in Swann's apparatus with a quadrant electrometer, and in Compton's apparatus with a Lindemann electrometer.

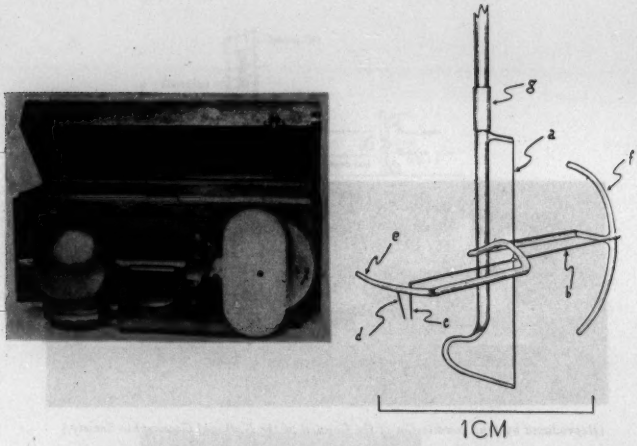


Figure 9. Apparatus by Millikan and Neher (1936). The sphere to the left contains the ionization chamber, capable of being surrounded by various thicknesscs of lead. The moving system is a quartz fibre assembly in which the motion of the arm *b* is controlled by the torsion of *a*. Everything from *g* downwards is coated with gold. The motion of *c* is recorded on the moving film of the camera and arrangements are made automatically to re-charge the system at fixed intervals.

(Reproduced, by kind permission, from the 'Physical Review'.)

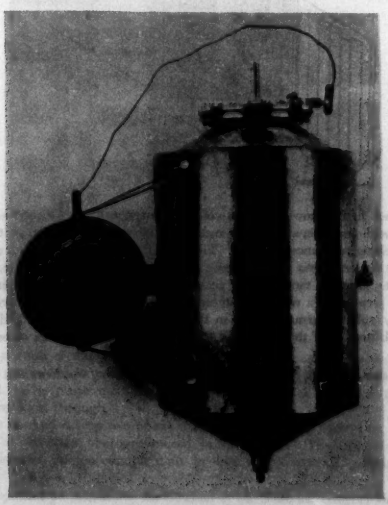
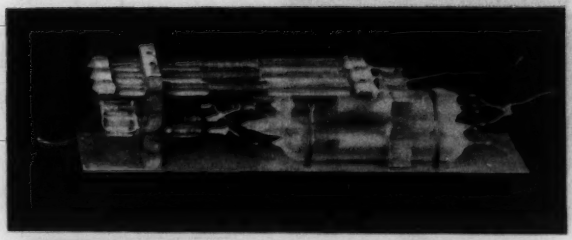


Figure 12. Apparatus by Bowen and Millikan (1933). Self-recording electrometer for observations in the stratosphere. Skylight falling through the 0·001-inch slit at the right casts shadows (diffraction images) of fibres on the photographic film placed just behind a horizontal slit in the recording attachment to the left.

(Reproduced, by kind permission, from the 'Physical Review'.)

To face page 28



(Reproduced by kind permission of the Council of the National Geographic Society.)

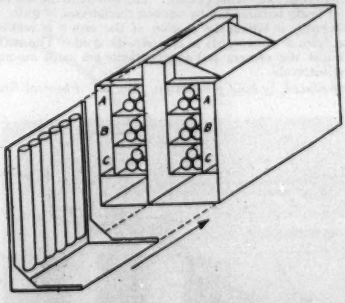
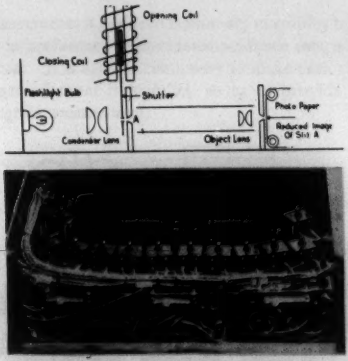
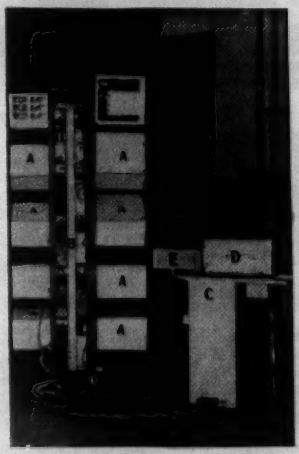


Figure 13. The upper figure shows a Geiger-counter unit, comprising three Geiger counters (the larger tubes) with their separate condensers (not very clearly shown) to the left, and their separate quenching resistances on top. This unit feeds to the grid of an amplifier.

The lower figure shows two Geiger-counter telescopes side by side. Each telescope comprises three units in tandem and each unit comprises three Geiger counters. The vertical tubes represent a wall of counters which can be slid into place as indicated by the arrow. By suitable arrangements in the electrical circuit, any record from one of the telescopes is obliterated if that record is accompanied by a discharge of one of the "wall" counters.

(Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.)





The optical recording mechanism.

Figure 14. The upper portion of the cut is a diagram of one of the optical recording units. A group of units is shown below. Normally the slit at A in the diagram was closed by the shutter behind it. The pulse operated the upper part of the solenoid, pulling up the plunger and letting light through.

The ten boxes marked A (top two open) contained cosmic-ray telescopes. The passage of cosmic rays was recorded photographically in camera D. The impulses initiated by the rays also registered on ten counter dials in box C. Camera D photographed these dials.

(Reproduced by kind permission of the Council of the National Geographic Society.)

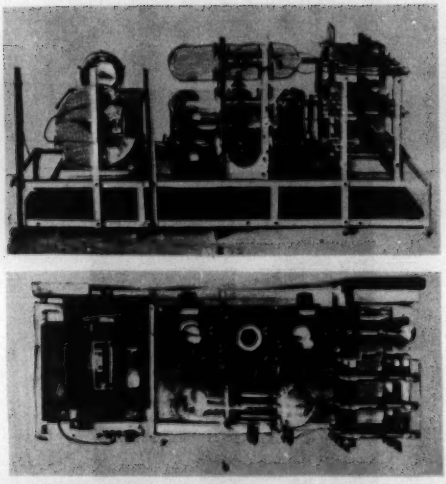


Figure 16. Plan and elevation of apparatus by Pfotzer (1936). The recording apparatus is to the left, the Geiger counters are in the centre, and the associated subsidiary apparatus is in the centre and to the right. The whole apparatus was enclosed in a cellophane shield to provide for constancy of temperature.

To face page 29

(iii) *Geiger counter telescopes*

For intensity measurements it has been customary to employ triple-coincidence sets, where possible, in preference to double-coincidence sets, so as to minimize accidental coincidences. It is further customary to make each of the three units comprise three Geiger counters in parallel, so as to provide better recovery conditions for an assigned counter area.

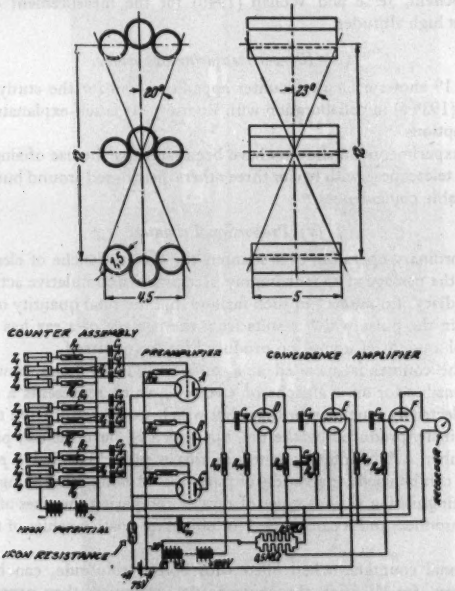


Figure 15. The upper figure shows a triple coincidence set, each element comprising three individual counters. The dimensions are in centimetres.

The lower figure shows the electrical circuit. Its meaning will be sufficiently clear from the information attached, together with the following scheme of magnitudes:—

$R_1 - R_9 = 10^9$ ohm, $C_1 - C_9 = 25\ \mu\mu$F.,
$R_{10} - R_{13} = 10^8$ ohm, $C_{10} = 100\ \mu\mu$F.,
$R_{15} = 0{\cdot}5 \times 10^8$ ohm, $C_{11} = C_{12} = 250\ \mu\mu$F.,
$R_{14} = R_{15} = R_{17} = 31 \times 0^8$ ohm, $C_{13} = 1000\ \mu\mu$F.,
$R_{16} = R_{18} = 5 \times 10^6$ ohm. $C_{14} = 1\ \mu$F.

In order to avoid spurious coincidences resulting from "showers", shielding counters have occasionally been employed in such fashion as to render the telescope "dead" to rays which pass through the shield and might operate each unit of the telescope simultaneously without having passed through the axis of the telescope. Figures 13 and 14 show an elaborate counter-telescope apparatus of this kind used by Swann, Locher and Danforth (1936a, 36b) and by Swann and

Danforth (1939) for both ground and high-altitude measurements in the gondola carried by a large balloon.

Figures 15 and 16 show a counter-telescope suitable for high-altitude work and designed by Pfotzer (1936) for photographic recording. Figure 17 shows an apparatus of the type used by Johnson (1937), in which the counts were recorded by radio on a ground station. Figure 18 shows a counter telescope apparatus used by Schein, Jesse and Wollan (1940) for the measurement of mesotron intensity at high altitudes.

(iv) *Shower-measuring apparatus.*

Figure 19 shows a Geiger counter apparatus used for the study of showers by Swann (1939 a) in collaboration with Ramsey. It is self-explanatory with the aid of its captions.

Many experiments on showers have been made by the use of single counters (or counter telescopes), with two or three others distributed around but arranged to record suitable coincidences.*

(v) *Proportional counters*

In the ordinary operation of a counter, the first avalanche of electrons associated with the passage of an ionizing ray precipitates a cumulative action through the intermediary of photons † in such fashion that the total quantity of electricity discharged in the pulse which results from the passage of a ray has no relation to the actual amount of ionization produced by the ray itself.

When the counter is operated at a sufficiently low voltage, however, each ray is responsible for an avalanche of electrons which represents a total transference of electricity much larger than that which would result from the number of ions originally produced by the ray, but which is, nevertheless, proportional to that number. A counter operated in this manner is called a *proportional counter*, and can be used for a variety of purposes. Among other things, it can be used for distinguishing various types of rays by comparing the sizes of the pulses which they produce, and it can be used for obtaining relative values of the specific ionization.

Proportional counters, when filled with boron trifluoride, can be used to count neutrons, for although the neutrons do not ionize, they promote disintegrations in the boron trifluoride with the resulting emission of ionizing radiations which can be recognized from other radiations by the sizes of the pulses which they give. Such counters have been used by Korff (1939, 41 b–41 d) and by Korff and Clarke (1942) to count neutrons at high altitudes.

(vi) *Cloud chambers*

Cloud chambers are frequently used in combination with Geiger counters in such a manner as to promote an expansion only when a cosmic ray passes through the apparatus. They are also frequently used in conjunction with a magnetic field so as to provide for energy measurements, and by counting droplets they can provide data on ionization per centimetre of path.

* See, for example, Woodward (1936), Nielsen (1938), Bothe (1939 a, 39 b).

† More detailed accounts of these matters, together with further references, will be found in papers by Korff (1939, 41 a), Rose and Korff (1941), Montgomery and Montgomery (1941 a, 41 b), Ramsey (1942), Ramsey and Hudspeth (1942), Rose and Ramsey (1942), Weisz (1942), Weisz and Ramsey (1942).

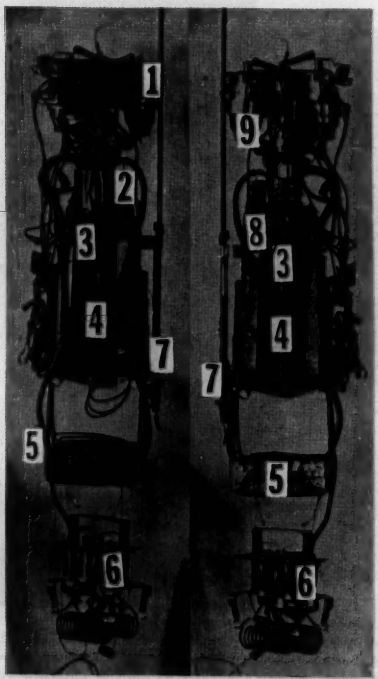


Figure 17. Front and rear views of the second model cosmic-ray radio transmitter. (1) Coincidence counters, (2) oil-cup containing contact points, (3) L.T. batteries, (4) H.T. battery, (5) high-voltage condensers, (6) radio transmitter, (7) barometer, (8) relay key connecting to the transmitter, (9) cosmic-ray counter circuit.

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

To face page 30

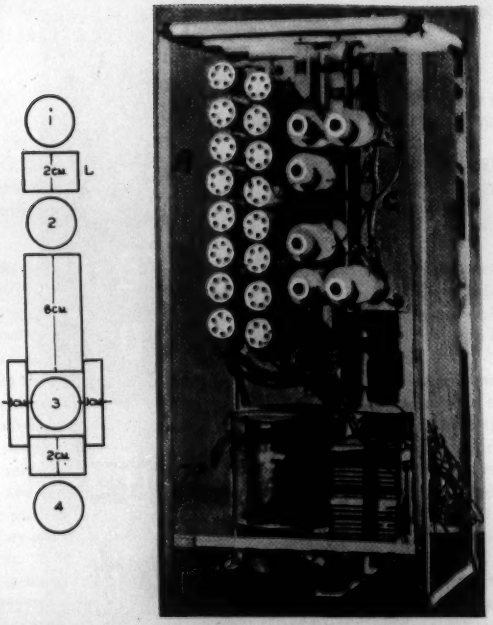


Figure 18. In the figure at the left, 1, 2, and 3 are counters arranged for three-fold coincidences; 2, 3 and 4 constitute another coincidence set. These coincidences are recorded photographically, and a coincidence for either set must be associated with the particle which is able to traverse at least 8 cm. of lead interposed in the form of blocks between the counters.

In the figure to the right, C represents the counters, A the amplifying tubes, and R the recorder.

(*Photograph reproduced, with acknowledgments, from 'Symposium sobre Raios Cosmicos'.*)

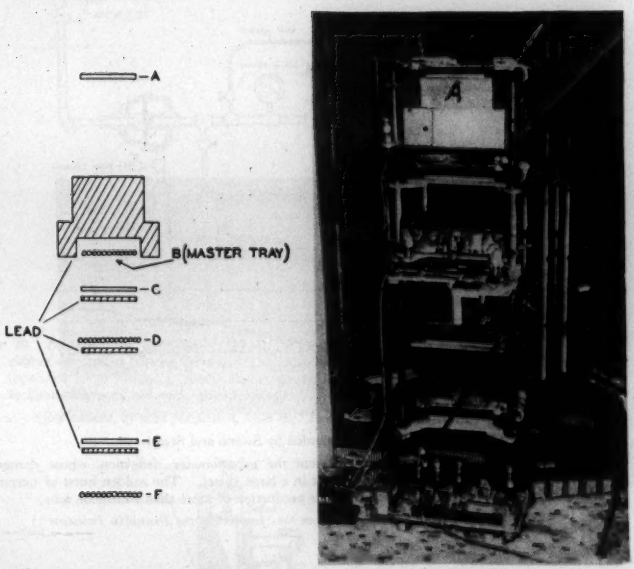
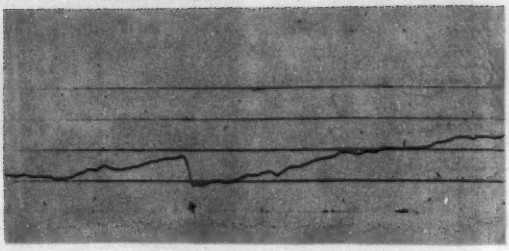


Figure 19. The apparatus is shown schematically at the left. A, B, C, D, E, F are counter trays. Each individual counter in each tray, except those of B, is connected to a separate electroscope, of which there are 180. Deflections of these electroscopes are recorded photographically. The electrical circuits are arranged so that none of the electroscopes are permitted to operate unless at least one counter in B, one in A, and one in F discharge simultaneously.

In the figure at the right, A indicates the large mass of lead. Immediately above and below this mass are the two counter trays (A and B, of the left-hand figure). To the right of each will be seen boxes, here labelled B, which contain electroscope systems. Halfway down the figure is another counter tray, quite visible, and below it, in succession, are three others together with another box, B, containing electroscopes.

(Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.)



A nuclear explosion recorded by Swann and Street.

Figure 23. The ordinates of the trace represent the galvanometer deflection, whose changes measure changes in the ionization current in a large vessel. The sudden burst of current indicated by the arrow corresponds to the production of more than 10 million ions.

(Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.)

To face page 31

For the study of shower phenomena, slabs of dense material such as lead are frequently incorporated inside the chamber ; also, cloud chambers have been operated at high pressures for the purpose of securing in the chamber itself a greater equivalent gas path and a greater time of sensitivity. Tracks obtained

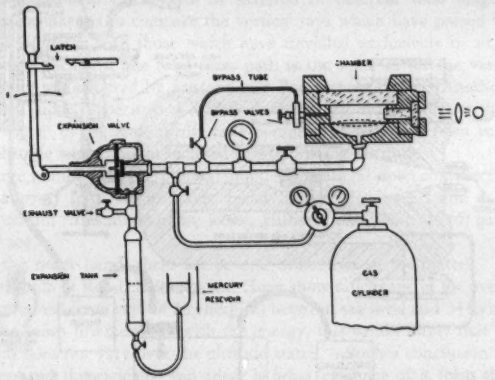


Figure 20. Apparatus by Mott-Smith (1934). The volume expansion is realized by a flexible rubber diaphragm (shown dotted). The expansion is promoted by an " expansion valve " (operated by a " latch ") which releases gas into the expansion tank. The by-pass is to facilitate filling or emptying the chamber.

(*Reproduced, by kind permission, from the ' Review of Scientific Instruments '.*)

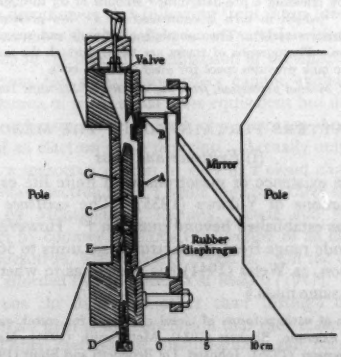


Figure 21. Apparatus by Blackett (1934). The piston A, which controls the expansion, is supported by the rubber diaphragm which closes the chamber. In the expanded condition it lies against the rubber-covered iron plate C. Three long screws D serve to adjust the amount of the expansion. A magnetic field is provided and the tracks are photographed by two cameras, using a mirror at 45°. The expansion valve is operated by Geiger counters suitably placed.

(*Reproduced, by kind permission, from the ' Proceedings of the Royal Society '.*)

at high pressures are exceptionally free from diffusion.* Figures 20, 21 and 22 show three types of cloud chambers.

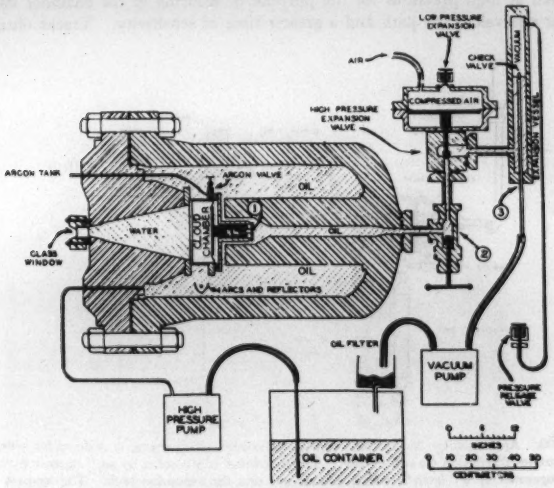


Figure 22. Hydrostatically-supported cloud chamber for operation at pressures up to 200 atmospheres by Johnson, De Benedetti and Shutt (1943). The pressure is sustained by the walls of the oil-filled steel tank rather than by the glass parts of the chamber itself. Expansions are made by releasing a pre-determined amount of oil through the "high-pressure expansion valve" (which in turn is controlled by a "low-pressure expansion valve") into the "expansion vessel". The complete oil circuit and some auxiliary and safety devices are shown. Photographs of tracks are taken through the small "glass window". The length of the tank provides space for magnetic-field coils.

(*Reproduced, by kind permission, from the 'Review of Scientific Instruments'.*)

§8. MATTERS PERTAINING TO THE MESOTRON

(i) *Existence and mass*

Evidence for the existence of mesotrons with finite life came first from the mathematical predictions of Yukawa (1935). The *existence* of the mesotron may be considered as established beyond question.† However, the mass determinations cover a wide range from 39 electron mass units to 569 units.‡ There is even some question, as Weisz (1941) has shown, as to whether all mesotrons have in actuality the same mass.§

* Many descriptions of various forms of cloud chambers for cosmic-ray investigations have been given. See, for example, Blackett (1934), Mott-Smith (1934), Herzog (1935), Locher (1936), Street and Stevenson (1936), Johnson, De Benedetti and Shutt (1943).

† See article by Street (1939).

‡ See article by Peierls (1939).

§ Recent articles summarizing information concerning the mesotron have been written by Peierls (1939), Schein and Wilson (1939 a), Rossi (1939), Jones (1939), Neddermeyer and Anderson (1939).

(ii) *Lifetime of the mesotron*

The experimental evidence for finite life of the mesotron is less complete than is that for the existence of the mesotron itself. It rests primarily upon a comparison of the intensity of the radiation after passing, for example, through two mass-equivalent thicknesses of material of different total lengths. One obvious procedure is to compare the vertical rays which have passed through a high tank of water with those which have travelled exclusively in air, but at a zenith angle such that the total mass path is the same as for the vertical rays. Such a plan was adopted by Johnson and Pomerantz (1939). Another method is founded upon experiments which secure equal mass paths with different time of transit in the same medium by comparing the results from vertical rays at one altitude with those of inclined rays at a higher altitude.

A large number of experiments, founded in general upon considerations such as the above,* have led to a rest mean lifetime between 2 and 4×10^{-6} sec. Recent careful measurements by Rossi, Hilberry and Hoag (1940) gave a value 2×10^{-6} sec.

On the other hand, there are several difficulties in the matter. Thus, the measurements of Rossi, Hilberry and Hoag show no change in the average mean life (as obtained from rays of all energies) between sea level and 3440 m. altitude. Since the mean life increases with the energy, this would imply that the energy spectrum does not vary over the altitude stated. Such a conclusion is not easy to understand theoretically, but there is other evidence of it from the work of Woodward and Street (1936). Moreover, there is other evidence for it in the form of variation of mesotron intensity with altitude, to be discussed later.

Another element results from the fact pointed out by Swann (1938 b) and by Fermi (1940) to the effect that the ionization loss in dense materials is reduced by the polarization of the material by the ionizing ray. This would have an effect in the same direction as that following from mean life considerations, and indeed, calculations show that the effect is comparable with those resulting from such considerations in the case of a comparison of the effects in gases and dense materials. However, the polarization effect cannot explain the difference in absorption by columns of air of equal mass equivalent but different extents.

Again, theory indicates that the death of a mesotron should be accompanied by the emission of an electron and a neutrino. Actually only five reliable photographs, in which a mesotron is seen to stop in a cloud chamber, are available. Two of these show an electron track,† while the other three show no such track.

Experiments by Montgomery, Ramsey, Cowie and Montgomery (1939), designed to detect by counter methods the ejection of an electron at the end of the range of the mesotron, have failed to reveal it. On the other hand, evidence of its existence is afforded by experiments of Rasetti (1941).

That mesotrons do disappear other than through death by ionization is much more certain than is the conclusion that they disappear by spontaneous disintegration. Other methods of disappearance can occur by the mesotron

* See Peierls (1939).

† The two showing a track were obtained by Williams and Roberts (1940) and by Shutt, De Benedetti and Johnson (1942), while those showing no track were obtained by Neddermeyer and Anderson (1938), by Nishina, Takeuchi and Ichimiya (1939), and by Johnson and Shutt (1942).

rtansferring its energy to a non-ionizing particle such as a photon, or a neutral mesotron (a neutretto).

(iii) *Other matters pertaining to mesotrons*

We have already referred to the possible existence of neutrettos. Evidence for the existence of such entities, or at any rate of penetrating non-ionizing entities, has been obtained by Schein and Wilson (1939 b), by Schein, Jesse and Wollan (1940), by Rossi and Regener (1940) and by Jánossy and Rochester (1941, 42). This evidence came from the observation of mesotrons produced by non-ionizing entities whose penetration was such as to indicate that they were not photons.

Detailed examination of the data from extensive showers seems to indicate * that such showers contain mesotron-shower particles. Evidence of the non-existence of such mesotron showers is, however, provided by experiments of Jánossy (1938). Recent experiments of Schmeisser (1938, 39), Bothe (1939 a, 39 b), Rochester (1944), Jánossy and Rochester (1942) claim to establish the existence of showers of mesotrons. However, these conclusions disagree with those of Nielsen, Morgan and Morgan (1939).

Swann and Ramsey (1940 a) used an apparatus, figure 19, in which there was an 18-cm. block of lead. Above this was a tray of 18 counters, and below it several similar trays with 1-cm. lead plates between. The apparatus was arranged to record only when at least one penetrating ray passed through all the trays. In 2,300 such events there were nine cases where at least 10 counters in the upper tray recorded and in which the rays passed through the 18-cm. block and discharged corresponding counters in the trays below. On the other hand, in 5,150 events only 47 cases were observed in which more than a single *mesotron* ray passed through the apparatus. These experiments were performed below a 30-ft. tank of water. Evidence for the production of secondary mesotrons by mesotrons has been provided by Maier-Leibnitz (1939).

(iv) *Theories on the creation of mesotrons*

The problem of the creation of mesotrons is one invoking the most recent aspects of nuclear physics, and is too involved for treatment here. For the underlying theory concerning such matters, one may consult the articles by Bethe (1940 a, 40 b). The creation of mesotrons by photons has been discussed by Bhabha (1938), Heitler (1938), Yukawa, Sakata, Kobayasi and Taketani (1938), Nordheim and Nordheim (1938), Kobayasi and Okayama (1939) and by Heisenberg (1939). The creation of mesotrons from neutrettos has been discussed by Arley and Heitler (1938). A comprehensive discussion of the production of mesotrons from protons is given by Hamilton, Heitler and Peng (1943), and a general theoretical summary of mesotron production has been given by Peierls (1939).

§9. ELECTRON-SHOWER PHENOMENA

(i) *Shower production by pair formation*

The fact that the total radiation could be increased by passage through a slab of metal was observed by Steinke (1927, 28), Schindler (1931), Hoerlin (1936),

* See Euler and Wergeland (1939), Auger, Maze and Robley (1939), and Auger, Maze, Ehrenfest and Fréon (1939).

and Street and Young (1934). For some time the details of such phenomena were rather obscure, but the development of the pair-formation theory finally provided, in the hands of Carlson and Oppenheimer (1937) and others, a consistent picture in which to understand them, and stimulated a large amount of research on both the experimental and theoretical sides. Observations of the creation of showers of rays as the result of a single ray passing through a slab of material have been made in general by three techniques—that of the cloud chamber* (usually counter-controlled), that involving the direct use of Geiger counters for observing the showers—as, for example, the apparatus shown in figure 19—and that involving ionization chambers.† The essential features of these phenomena are comprised in the statement that a primary high-energy electronic ray entering a slab of lead, for example, will, in general, give rise—by pair formation—to electrons and photons from the emergent side, the number of such electrons and photons increasing with the thickness of the lead up to a certain maximum and then diminishing.

(ii) *Bursts*

Occasionally such a large number of secondary rays accompany a single primary event as to place the phenomenon in a category different from that usually designated under showers, and in such cases the phenomena are referred to as bursts. They were first observed by Hoffmann and Lindholm (1928) and by Hoffmann and Pforte (1930). A large burst of this kind obtained by Swann and Street is shown in figure 23 ‡. At first it was customary to attribute such a burst to something of the nature of an atomic explosion. However, the present tendency is to believe that they can be accounted for on the same lines as the ordinary showers, but originated by a parent (photon or electron) of very high energy.§

(iii) *Electron showers produced by mesotrons—showers below sea level*

Mesotrons, by a process closely related to ionization, can eject an electron from an atom. The electron can then give rise to a shower by pair formation. The theory of this matter has been given by Bhabha (1938). Swann and Ramsey (1939, 40 b), using the apparatus shown in figure 19, found that the number of cases where different numbers of electron rays accompanied a mesotron were in general agreement with Bhabha's predictions, except that the number of cases in which a mesotron is accompanied by a single electron ray decreases with the atomic number of the material used instead of being practically independent of it, as indicated by Bhabha's theory. Using another form of apparatus, Swann and Ramsey (1938) demonstrated the existence of electron showers in a mine ‖

* For cloud-chamber data see, for example, Blackett and Occhialini (1933), Anderson (1933), Anderson and Neddermeyer (1936), Fussell (1937), and Street (1939).

† For Geiger-counter observations and ionization-chamber observations, see, for example, Rossi (1933), Swann and Montgomery (1933 a, 33 b), Rossi and De Benedetti (1934), Gilbert (1934), Woodward and Street (1935), Geiger and Zeiller (1935), Johnson (1935), Priebsch (1935), Montgomery and Montgomery (1935, 36), Woodward (1936), Morgan and Nielsen (1937), Nielsen (1938), Swann and Ramsey (1938), Wilson (1938 a, 38 b) and Swann (1939 a).

‡ Facing page 31.

§ See, for example, Schein and Gill (1939), where much additional information concerning these bursts is given, together with references. See also Nielsen (1938).

‖ Recent observations along related lines have been made underground by Ehmert (1937) Wilson (1938 a, 38 b), and by Nielsen (1938) in collaboration with Morgan.

100 feet deep and showed that, on the average, each mesotron ray was accompanied by 1·6 electron rays of energy about 2×10^8 ev.

(iv) *Extensive showers*

A specially dramatic form of shower is one where particles, known as corresponding to a single shower by the simultaneity of their occurrence, are observed at wide distances apart. Phenomena of this kind, with simultaneity of occurrence of particles at a distance of 1·5 m., were observed by Swann (1938 a, 38 c) in collaboration with Ramsey. At about the same time, Auger, Maze and Grivet-Meyer (1938), Auger and Maze (1938), Auger, Maze, Ehrenfest and Fréon (1939) observed similar phenomena, which became extended by Auger, Maze and Robley (1939) in subsequent investigations to the observation of simultaneous shower particles at distances up to 300 m. It is now customary to attribute such phenomena to showers produced in the stratosphere by primary particles of exceedingly high energy, the offspring of these events being observable at sea level. Calculations of the energy of the primary particles necessary to account for such showers lead to energies as high as 10^{16} ev.

(v) *Further phenomena pertaining to showers—Rossi curves*

If the intensity of the shower radiation is measured below a piece of lead, and if the intensity be plotted against the thickness, a curve will be obtained which first rises to a maximum and diminishes with corresponding rapidity, until at a distance comparable with 4 or 5 cm. it attains a condition in which it varies slowly and approximately linearly with the thickness. Such curves are called Rossi (1933) curves. Ackemann (1934) and Hummel (1934) observed such curves with the additional feature of a second hump in the neighbourhood of 15 cm. The existence of this hump has been questioned. Thus Schwegler (1936), Morgan and Nielsen (1937), Auger, Maze, Ehrenfest and Fréon (1939), and Altmann, Walker and Hess (1940) failed to find it. On the other hand, it has been observed by Drigo (1934, 35), Priebsch (1936), Clay, van Gemert and Wiersma (1936), Woodward (1936), Schmeisser and Bothe (1938), Swann and Ramsey (1940 c), de Vos (1940), and Broussard and Graves (1941); also Nielsen, Morgan and Morgan (1939) found the hump with one counter arrangement and failed to find it with another arrangement.

In the case of the experiments of Swann and Ramsey, the second hump was found in the case of both lead and tin, but at different thicknesses. The first hump is very readily explained as resulting from atmospheric electrons which, by pair formation in the dense material, give rise to a shower intensity-thickness curve of the same kind as that already described in connection with the maximum found for cosmic-ray intensity at high altitudes. After a thickness has been attained at which this phenomenon has died out, one would expect to observe only the showers representative of the electrons which had been knocked out of atoms by mesotrons, and electrons which had resulted from decay of mesotrons. These might, therefore, be expected to diminish with thickness in proportion to the absorption of the mesotrons. The second hump is not so readily explained. If one examines the curves of Swann and Ramsey (1940 c) for tin, it will be seen that the second hump has features similar to the first, features, moreover, asso-

ciated with the behaviour of electrons which could only have originated in connection with some nuclear process taking place in the vicinity of the said hump, a process concerned with the creation or annihilation of mesotrons, for example.*

(vi) *Neutrons in the cosmic radiation*

Search for neutrons in the cosmic radiation has been made by several investigators. A summary of the essential results is presented in figure 24, which gives results from various observers and compares the data with those for the large "bursts". The article by Korff (1939), from which this graph is taken, gives many references to the work of former investigators.

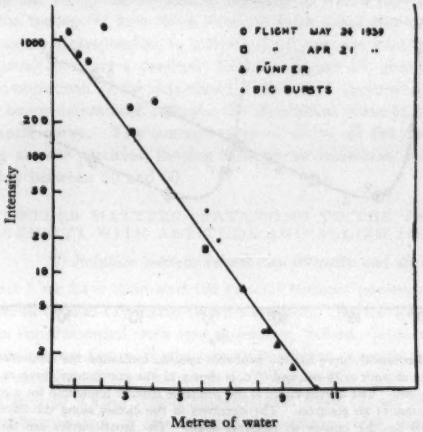


Figure 24.
(*Reproduced, by kind permission, from the 'Reviews of Modern Physics'.*)

A few pertinent facts relating to cosmic-ray neutrons are the following : According to the measurements of the Montgomerys (1939), the neutron intensity at sea level is about one-sixteenth of the cosmic-ray intensity. However, Korff finds that, from sea level to an altitude comprising nine-tenths of the atmosphere, the neutron intensity increases by a factor of 650, which is very much more than the corresponding increase for the cosmic-ray intensity. Korff (1939) gives reasons for supposing that neutrons do not exist in the primary radiation. Hamilton, Heitler and Peng (1943) regard them as formed from primary protons during processes associated with the creation of mesotrons. According to Bethe, Korff and Placzek (1940), the life history of an originally fast neutron in the atmosphere is as follows : The neutron first loses energy in large amounts by collisions with nitrogen atoms, with the emission of nuclear particles. After it has been reduced in energy in this manner to about 4×10^8 ev., further loss by nuclear processes ceases to occur and the only losses are such collisional losses as

* Chandrashekhar Aiya (1944) has recently reported a sudden drop in mesotron intensity after passing through 21 cm. of lead.

ordinary atoms of low energy experience, and which ultimately result in the neutrons attaining on the average an energy equal to the average energy of the surrounding gas molecules.

§10. SPECIFIC IONIZATION* FOR HIGH-ENERGY PARTICLES

Ionization by charged particles is a rather complex phenomenon in which the numbers emitted per centimetre of path decrease with the increase of the energy with which they are emitted. Although most of the ions emitted have practically zero energy, those emitted with higher energies also ionize, and the smallness of their numbers is partially compensated by the largeness of their

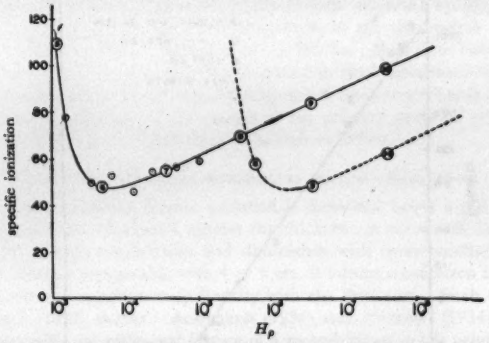


Figure 25. The theoretical curve for the probable specific ionization for electrons in air in ion pairs per cm. of track at 76 cm. and 0° C. is shown in the continuous curve as a function of $H\rho$ in gauss cm. The dotted curve is the probable specific ionization for a particle of mass 250 × the mass of an electron. The numbers in the circles along the curves indicate the power of 10 for the energy in electron volts. The small circles are the experimental observations of Corson and Brode.

(*Reproduced, by kind permission, from the 'Reviews of Modern Physics'*)

energy to the end of making them responsible for an appreciable fraction of the total number of ions produced.

The theory of ionization was developed on the classical theory by Thomson (1912) and extended to include relativity and other considerations by Bohr (1915). The implications of the relativistically classical theories for very high energies of the ionizing particle have also been discussed by Swann (1924). The Thomson theory, even when presented relativistically, gives a picture in which the specific ionization decreases with increase of velocity of the ionizing particle and attains a sensibly constant value as the velocity of light is approached. Quantum theories, and, indeed, Bohr's extension of the original classical theory, give a result in which the specific ionization decreases with increase of the energy of the particle, attains a minimum, and then increases with further increase of energy.

Quantum theories of ionization have been developed by Williams (1932), Møller (1932), Bethe (1933), Bloch (1933), and Oppenheimer (1935) and when

* Ionization per centimetre of path.

a limit is set to the maximum energy lost to a secondary electron there is general agreement that the total number N of pairs of ions per centimetre produced by all secondaries below this limit is given by an expression of the form

$$N = \frac{A}{\beta^2}\left(\log k + \log \frac{\beta^2}{1-\beta^2} - \beta^2\right), \qquad \ldots\ldots (5)$$

where β is the ratio of the velocity of the particle to that of light, while k and A are constants for a given kind of particle.

Experimental data have been obtained by methods in which the integrated ionization from a particle travelling a certain distance has been obtained by measuring the integrated ionization currents to which they give rise, but data suitable for testing (5) have been obtained from cloud-chamber observations in which droplets corresponding to individual ion pairs are counted and ray energies are measured by using a magnetic field.* Figure 25, given by Brode (1939), shows a comparison of the experimental data with expression (5), the constants A and k being determined to make the theoretical curve fit two features of the experimental curve. The average value of N for all the cosmic-ray particles, including shower particles, passing through an ionization chamber at sea level is probably between 50 and 60.

§11. FURTHER MATTERS PERTAINING TO THE VARIATION OF INTENSITY WITH ALTITUDE AND ALLIED PHENOMENA

(i) *Relation between cosmic-ray intensity and altitude*

In Part I we have discussed the general features pertaining to the variation of the overall cosmic-ray intensity with altitude. By the use of Geiger-counter telescopes supplemented with lead absorbers, Schein, Jesse and Wollan (1940) have measured, see figure 18, the vertical mesotron intensity up to high altitudes and have established the existence of a maximum. For altitudes extending from sea level to 4 km., Rossi, Hilberry and Hoag (1940) have made specially accurate observations of the mesotron intensity with the results shown in figure 26, in which the logarithm of the intensity is plotted against the depth. A constant absorption coefficient would demand a linear relation which does not correspond to figure 26. It is interesting, however, to examine these results from another standpoint. If we plot the logarithm of the intensity against actual altitude in kilometres, we obtain figure 27, showing a very exact linear relationship and corresponding, therefore, to an exponential increase with altitude and to a constant coefficient $\mu_{\text{atm}} = 0{\cdot}205$. We may thus regard it as an experimental fact that, if I_{m} is the mesotron intensity, and x the distance above sea level in kilometres

$$I_{\text{m}} = I_{\text{m}0}\, e^{0{\cdot}205x}, \qquad \ldots\ldots (6)$$

where $I_{\text{m}0}$ is the mesotron intensity at sea level.

Now if h is the water-equivalent depth below the top of the atmosphere, and h_0 the water-equivalent depth of the atmosphere,

$$h = h_0 e^{-\lambda x}, \qquad \ldots\ldots (7)$$

where $\lambda = 0{\cdot}120$ if x is in kilometres.†

* The article by Brode (1939) contains many references to measurements of the specific ionization by these methods.

† We have taken this value of λ as the value used by Rossi, Hilberry and Hoag (1940).

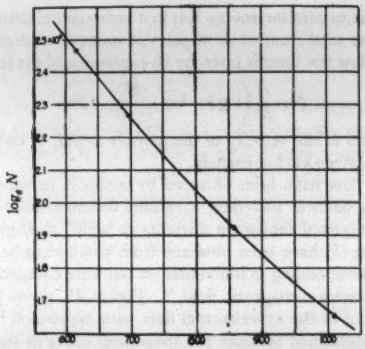


Figure 26. Logarithm of intensity of cosmic-ray mesotrons as a function of the depth. To express the results in terms of water-equivalent altitude, the abscissae should be read from right to left.

(*Reproduced, by kind permission, from the 'Physical Review'.*)

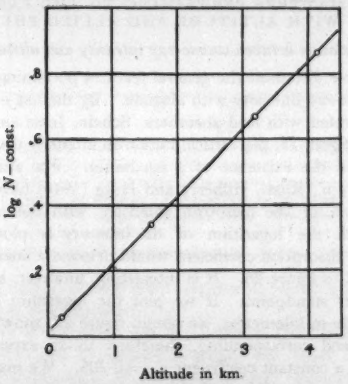


Figure 27. Logarithm of intensity of cosmic-ray mesotrons as a function of altitude in kilometres.

Equations (6) and (7) yield

$$\frac{I_m}{I_{m0}} = \left(\frac{h_0}{h}\right)^n \qquad \ldots\ldots (8)$$

if $\lambda n = 0{\cdot}205$, i.e. $n = 1{\cdot}71$.

Thus the statement that I_m is exponential in true altitude x is the same as the statement that I_m is of the form A/h^n. The latter form has been accepted for some time in relation to the so-called "hard component" of the cosmic radiation, even before mesotrons were known. However, it seems to have been customary

to regard it as an approximation to an exponential in h, with an absorption coefficient varying with h. The simplicity of the story in the form of a true exponential law, when expressed in terms of true distance, does not seem to have been observed.*

Now it is well known† that an intensity law of the form $I_m = A/h^n$ is the type yielded by a pure ionization absorption, with a constant range for each energy, as applied to a spectral-energy distribution law of the type‡ $B/(E+\gamma h)^{n+1}$, where γ is the ionization energy loss per unit of water-equivalent path. It might be thought that the form $I_m = A/h^n$ *demands* these conditions. However, such is not the case. Thus, an energy-distribution law of the form

$$F(E) = [f(Eh^n)],$$

where f is an arbitrary function, yields

$$I_m = \int_0^\infty f(Eh^n)dE = \frac{1}{h^n}\int_0^\infty f(\zeta)d\zeta,$$

where the integral does not involve h. Again, a law of the form $F(E) = f(E)/h^n$, where f is an arbitrary function, obviously satisfies these conditions.

Thus, the fact that the law of the form $I_m = A/h^n$ comes naturally from a range phenomenon as applied to a spectral distribution law of the form $B/(E+\gamma h)^{n+1}$ does not preclude the possibility of its also coming from more elaborate mechanisms involving, in addition to ionization, mean life considerations, for example.

(ii) *The law for total intensity*

The kinetic energy of all the rays, mesotrons, electrons, etc., ultimately becomes dissipated in ionization.§ In addition, a certain fraction‖ (say η) of the rest energy of the mesotrons becomes converted into electronic energy, and so ultimately into ionization energy. The dissipation per kilometre on this account is $m_0c^2\eta dI_m/dx$, where x is in actual distance and is measured upwards. If I_t is the total intensity (mesotrons plus electrons), the ionization loss per kilometre is $I_t\alpha_0 e^{-\lambda x}$, where α_0 is the ionization per kilometre at sea level, and $\alpha_0 e^{-\lambda x}$ is the corresponding value at the altitude x.

Now if $F(E, x)$ is the spectral distribution function for the mesotrons at the altitude x, so that

$$I_m = \int_0^\infty F(E, x)dE,$$

the energy provided per second by disintegration in the group $F(E, x)dE$ is $(1-\beta^2)^{\frac{1}{2}}EF(E, x)dE/\tau_0$, where τ_0 is the rest mean life and β is the ratio of the velocity to the light velocity c. Thus the energy provided per kilometre by this group is $(1-\beta^2)^{\frac{1}{2}}EF(E, x)dE/\beta c\tau_0$. This is equal to $m_0EF(E, x)dE/\rho\tau_0$, where

* Incidentally, the value of n is in satisfactory agreement with the value 1·93 cited by Clay (1939) from experiments of another kind.

† See, for example, Swann (1938 c) for detailed theory.

‡ This corresponds, of course, to a law of the form B/E^{n+1} for the incoming particle with, of course, a lower limit determined by the magnetic field.

§ Unless, of course, some of the rays are transformed into non-ionizing entities such as neutrettos, and provided that such entities, whatever they may be, do not suffer reconversion into ionizing entities.

‖ The remainder of the rest mass going into energy of neutrinos.

p is the mesotron momentum corresponding to the velocity βc. Thus, combining our results,

$$I_t\alpha_0\epsilon^{-\lambda x} = m_0c^2\eta\frac{dI_m}{dx} + \frac{m_0}{\tau_0}\int_0^\infty \frac{E}{p}F(E, x)dE + \lambda I_m\alpha_0\epsilon^{-\lambda x}. \quad \ldots\ldots (9)$$

Now it is easy to show that *

$$cp = (E^2 + 2m_0c^2E)^{\frac{1}{2}},$$

so that

$$\frac{E}{p} = \frac{c}{\left(1+\frac{2m_0c^2}{E}\right)^{\frac{1}{2}}}.$$

For high energies $E/p = c$, and if the spectral distribution curve is such that there is only small contribution to I_m from rays whose energy is not considerably greater than m_0c^2, we could write $E/p = c$ in the integral in (9), in which case the term involving this integral would become m_0cI_m/τ_0. Since, moreover, experiment has shown I_m to be exponential in x with a coefficient $\mu_{sm} = 0{\cdot}205$, (9) yields

$$I_t - I_m = I_{m0}\frac{m_0c}{\alpha_0}\left(c\eta\mu_{sm} + \frac{1}{\tau_0}\right)e^{\lambda x}e^{\mu_{sm}x}. \quad \ldots\ldots (12)$$

Under these conditions $I_t - I_m$ would be exponential in x and would have the coefficient $\mu_{sm} + \lambda = 0{\cdot}205 + 0{\cdot}120 = 0{\cdot}325$. Such a culmination of the relationship of the coefficients for the mesotron intensity and for the soft component is of such an ideal simplicity that one could but wish that it could be true. As the soft component becomes dominant, I_t approximates to an exponential with coefficient 0·325. Figure 28 shows the results of Greisen (1942) and figure 29 those of Swann and Danforth (1939) for the variation of total intensity with altitude as measured by coincidence counter sets. Figure 30 shows the intensities plotted logarithmically against altitude in kilometres. It will be seen that, except in the low altitudes, the graph corresponding to the observations of Swann and Danforth approximates to linearity up to about 6-km. altitude, corresponding over that range to a law exponential in actual altitude in kilometres. However, the coefficient is 0·4 instead of 0·325. This fact is probably symbolic of the situation already cited in Part I to the effect that mesotron decay alone is insufficient to provide for the soft component and that "extensive electron showers" must be regarded as playing a significant part.

In conclusion, it cannot be said that the story of the variation of intensity with altitude is complete, even for altitudes so low that we do not encounter complications arising from the maxima in mesotron intensity and electron

* We have

$$\frac{E}{m_0c^2} + 1 = \frac{1}{(1-\beta^2)^{\frac{1}{2}}} = \frac{p}{m_0\beta c}. \quad \ldots\ldots (10)$$

Writing

$$E/m_0c^2 \equiv \epsilon,$$
$$1-\beta^2 = (1+\epsilon)^{-2}$$
$$\beta = [1-(1+\epsilon)^{-2}]^{\frac{1}{2}}. \quad \ldots\ldots (11)$$

From (10) and (11)

$$p = m_0c\beta(1+\epsilon) = m_0c(\epsilon^2+2\epsilon)^{\frac{1}{2}}$$
$$cp = (E^2+2m_0c^2E)^{\frac{1}{2}}.$$

intensity. It will probably not become complete until we have a further understanding of the facts attending such matters as the transformation of mesotrons and neutrettos, and allied phenomena.

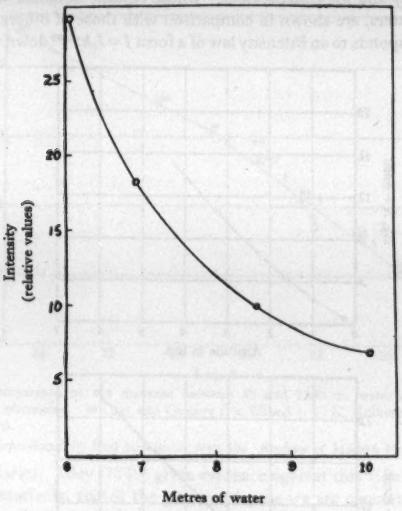


Figure 28. Total intensity, as obtained by Greisen (1942), plotted against water-equivalent depth.

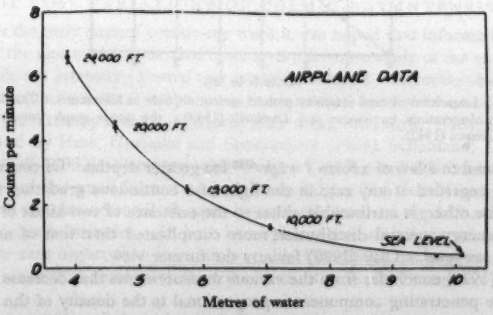


Figure 29. Total intensity, by Swann and Danforth (1939), plotted against water-equivalent depth.

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

(iii) *Further facts concerning the intensity below sea level*

Several observers have investigated cosmic-ray intensity at different depths below sea level in water, rocks, etc. Some of the most extensive data have been obtained by Clay and his collaborators,* whose results, obtained with threefold coincidence counters, are shown in comparison with those of others by figure 31. The graph corresponds to an intensity law of a form $I = I_0 h^{-1 \cdot 93}$ down to 311 metres

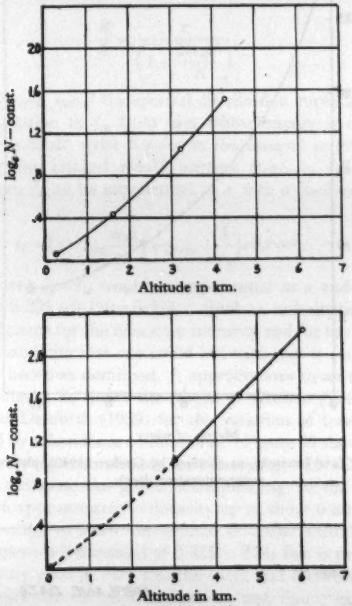


Figure 30. Logarithm of total intensity plotted against altitude in kilometres. The lower graph from observations by Swann and Danforth (1939); the upper graph from observations by Greisen (1942).

of water, and to a law of a form $I = I_0 h^{-3 \cdot 93}$ for greater depths. Of course, such a situation, regarded at any rate in the light of a continuous gradation from one form to the other, is attributable either to the existence of two kinds of particles or to an energy spectral distribution more complicated than that of an inverse power at sea level. Clay (1939) favours the former view.

Clay (1939) concludes from the various measurements that decrease of intensity of the penetrating component is proportional to the density of the material,

* See, for example, Clay (1927, 34), Clay, Wiersma and Grauff (1934), Clay, 't Hooft, Dey and Wiersma (1937), Wilson (1938 a), Swann and Ramsey (1938). Swann (1938 a, 38 b), Clay, Gemert and Clay (1939), Clay and Gemert (1939).

but is governed more by electron density and mean binding energy of the electrons.

It has been suggested by Barnóthy and Forró (1937) and by Wilson (1938 b) that the primaries under thick layers may be neutrinos, the observable entities

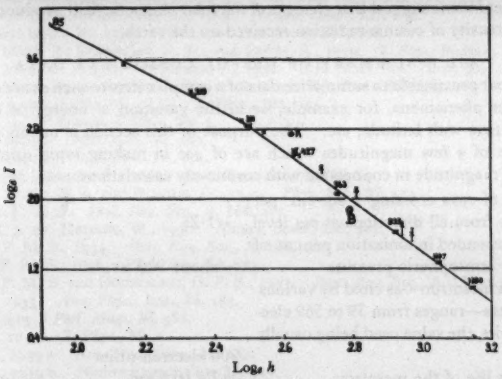


Figure 31. Comparison of the decrease between 85 and 1380 m. water equivalent found by different observers. ● Clay and Gemert ; x Wilson ; ⊙ K. Kolhörster ; ⊙ B. Barnóthy and Forró.

(*Reproduced, by kind permission, from the 'Reviews of Modern Physics'.*)

being secondaries. Clay (1939) gives evidence against this view. He concludes, moreover, tentatively, that at the greatest depths we are concerned with primary protons, a conclusion which would seem to call for careful examination in the light of theoretical considerations.

§ 12. TIME VARIATIONS OF COSMIC-RAY INTENSITY.

From the early days of cosmic-ray work it was hoped that information on the origin of the rays could be derived from a very accurate study of the variations of the cosmic-ray intensity (diurnal and seasonal changes). Accurate observations, after reduction to normal temperature and pressure, have shown a slight maximum of the intensity at noon, or shortly after noon, local time. Such results have been found by Hess, Grazadei and Steinmaurer (1934), Schonland, Delatinzky and Gaskell (1937), Compton and Turner (1937), and Thompson (1938). The cause of this small fluctuation (amplitude 0·2 per cent) is still unexplained.

There are also seasonal changes beyond the knee of the latitude curve at higher magnetic latitudes. In both hemispheres the intensity in winter is about 1 to 2 per cent higher than in summer. Blackett (1938) has suggested that the changes result from seasonal shifts in the mean altitude of the mesotron-producing layer.

The possibility of the existence of a variation of cosmic-ray intensity with sidereal time, discussed in Part I, is still not definitely rejected. In spite of an enormous accumulation of data by several observers, the results are indecisive : the effect, if it exists, is smaller than $\pm 0{\cdot}05$ per cent of the mean value.

Another interesting finding is the periodicity of cosmic-ray intensity correlated with the rotation of the sun (mean period 27 days). Hess (1936) and Graziadei (1936) found that small variations correlated with solar activity recurred within 27 to 28 days, and some of them could be followed up through many solar rotations. Such phenomena suggest that changes of the solar magnetic field produce changes in the intensity of cosmic radiation received on the earth.

§13. A SUMMARY OF USEFUL COSMIC-RAY DATA

It is not practicable to summarize data of a complex nature such as are involved in shower phenomena, for example, or in the variation of energy of different kinds of rays with latitude, etc. The purpose of this section is merely to cite a collection of a few magnitudes which are of use in making rapid estimates of orders of magnitude in connection with cosmic-ray calculations :—

Number of rays crossing 1 sq. cm. per minute from all directions at sea level	$=1{\cdot}25$
Energy expended in ionization per cm. of path at atmospheric pressure	=about 900 ev.
Mass of a mesotron—as cited by various observers—ranges from 39 to 569 electron units, the value used being usually about	200 electron units
Mean rest life of the mesotron	$=2\times10^{-6}$ sec.
Energy for an electron to enter vertically	$=1{\cdot}5\times10^{10}$ ev. at the equator $=0{\cdot}70\times10^{10}$ ev. at 40° magnetic latitude
Energy for a proton to enter vertically	$=1{\cdot}4\times10^{10}$ ev. at the equator $=0{\cdot}61\times10^{10}$ ev. at 40° magnetic latitude
Ratio of mesotron vertical intensity at the maximum to that at sea level	=11 at magnetic latitude 53°
Maximum of mesotron vertical intensity occurs at	14·2 kilometres at magnetic latitude 53°
Ratio of total vertical intensity at the maximum to that at sea level	=51 at magnetic latitude 52° =about 16 at the magnetic equator
Maximum of total vertical intensity occurs at	18 kilometres at magnetic latitude 52°
Coefficient of absorption of mesotrons at sea level	$=1{\cdot}47\times10^{-3}$ cm^2/gm.

REFERENCES

Ackemann, M., 1934. *Naturwissenschaften*, **22**, 169.
Aiya, S. V. C., 1944. *Nature, Lond.*, **153**, 375.
Alfvén, H., 1936. *Nature, Lond.*, **138**, 761.
Alfvén, H., 1938. *Phys. Rev.*, **54**, 97.
Alfvén, H., 1939. *Phys. Rev.*, **56**, 1060.
Altmann, G. O., Walker, H. N. and Hess, V. F., 1940. *Phys. Rev.*, **58**, 1011.
Alvarez, L. and Compton, A. H., 1933. *Phys. Rev.*, **43**, 835.

ANDERSON, C. D., 1933. *Phys. Rev.*, **43**, 368.
ANDERSON, C. D. and NEDDERMEYER, S. H., 1936. *Phys. Rev.*, **50**, 263.
ARLEY, N. and HEITLER, W., 1938. *Nature, Lond.*, **142**, 158.
ASTON, F. W., 1927. *Proc. Roy. Soc.*, A, **115**, 487.
AUGER, P. and LEPRINCE-RINGUET, L., 1933. *C.R. Acad. Sci., Paris*, **197**, 1242.
AUGER, P. and LEPRINCE-RINGUET, L., 1934. *Nature, Lond.*, **133**, 138.
AUGER, P. and MAZE, R., 1938. *C.R. Acad. Sci., Paris*, **207**, 228.
AUGER, P., MAZE, R., EHRENFEST, P., Jr., and FRÉON, A., 1939. *J. Phys. Radium*, **10**, 39.
AUGER, P., MAZE, R. and GRIVET-MEYER, T., 1938. *C.R. Acad. Sci., Paris*, **206**, 1721.
AUGER, P., MAZE, R. and ROBLEY, 1939. *C.R. Acad. Sci., Paris*, **208**, 1641.
BARNÓTHY, J. and FORRÓ, M., 1937. *Z. Phys.*, **104**, 744.
BETHE, H., 1933. *Handb. d. Phys.*, **24**, 1.
BETHE, H., 1940 a. *Phys. Rev.*, **57**, 260.
BETHE, H., 1940 b. *Phys. Rev.*, **57**, 390.
BETHE, H. and HEITLER, W., 1934. *Proc. Roy. Soc.*, A, **146**, 83.
BETHE, H., KORFF, S. A. and PLACZEK, G., 1940. *Phys. Rev.*, **57**, 573.
BHABHA, H. J., 1938. *Proc. Roy. Soc.*, A, **164**, 257.
BHABHA, H. J. and HEITLER, W., 1936. *Nature, Lond.*, **138**, 401.
BLACKETT, P. M. S., 1934. *Proc. Roy. Soc.*, A, **146**, 281.
BLACKETT, P. M. S., 1938. *Phys. Rev.*, **54**, 973.
BLACKETT, P. M. S. and OCCHIALINI, G. P. S., 1933. *Proc. Roy. Soc.*, A, **139**, 699.
BLOCH, F., 1933. *Ann. Phys., Lpz.*, **16**, 285.
BOHR, N., 1915. *Phil. Mag.*, **30**, 581.
BOTHE, W. 1930. *Z. Phys.*, **59**, 1.
BOTHE, W., 1939 a. *Rev. Mod. Phys.*, **11**, 282.
BOTHE, W., 1939 b. *Naturwissenschaften*, **27**, 305.
BOTHE, W. and KOLHÖRSTER, W., 1929. *Z. Phys.*, **56**, 751.
BOTHE, W. and KOLHÖRSTER, W., 1930. *Berl. Bericht.*, **26**, 450.
BOWEN, I. S. and MILLIKAN, R. A., 1933. *Phys. Rev.*, **43**, 695.
BOWEN, I. S., MILLIKAN, R. A. and NEHER, H. V., 1933. *Phys. Rev.*, **44**, 246.
BOWEN, I. S., MILLIKAN, R. A. and NEHER, H. V., 1938. *Phys. Rev.*, **53**, 855.
BRODE, R. B., 1939. *Rev. Mod. Phys.*, **11**, 222.
BROUSSARD, L. and GRAVES, A. C., 1941. *Phys. Rev.*, **60**, 413.
CARLSON, J. F. and OPPENHEIMER, J. R., 1937. *Phys. Rev.*, **51**, 220.
CLARKE, E. T. and KORFF, S. A., 1941 a. *J. Franklin Inst.*, **232**, 217.
CLARKE, E. T. and KORFF, S. A., 1941 b. *J. Franklin Inst.*, **232**, 339.
CLAY, J., 1927. *Proc. K. Akad. Amst.*, **30**, 1115.
CLAY, J., 1928. *Proc. K. Akad. Amst.*, **31**, 1091.
CLAY, J., 1930. *Proc. K. Akad. Amst.*, **33**, 711.
CLAY, J., 1932. *Proc. K. Akad. Amst.*, **35**, 1282.
CLAY, J., 1934. *Physica*, **1**, 363.
CLAY, J., 1939. *Rev. Mod. Phys.*, **11**, 128.
CLAY, J., BRUINS, E. M. and WIERSMA, J. T., 1936. *Physica*, **3**, 746.
CLAY, J. and VAN GEMERT, A., 1939. *Physica*, **6**, 497.
CLAY, J., VAN GEMERT, A. and WIERSMA, J. T., 1936. *Physica*, **3**, 627.
CLAY, J., 'tHOOFT, C. G., DEY, L. J. L. and WIERSMA, J. T., 1937. *Physica*, **4**, 121.
CLAY, J., WIERSMA, J. T. and GRAAFF, G. N., 1934. *Physica*, **1**, 659.
CLAY, P. H., VAN GEMERT, A. and CLAY, J., 1930. *Physica*, **6**, 184.
COMPTON, A. H., 1932 a. *Phys. Rev.*, **41**, 111.
COMPTON, A. H., 1932 b. *Phys. Rev.*, **41**, 681.
COMPTON, A. H., 1933 a. *Phys. Rev.*, **43**, 387.
COMPTON, A. H., 1933 b. *Trans. Amer. Geophys. Un.*, 154.
COMPTON, A. H., 1933 c. *Scient. Month.*, **36**, 75.
COMPTON, A. H., 1936 a. *Rev. Sci. Instrum.*, **7**, 71.
COMPTON, A. H., 1936 b. *Phys. Rev.*, **50**, 1119.
COMPTON, A. H., 1939. *J. Franklin Inst.*, **227**, 607.
COMPTON, A. H. and BETHE, H., 1934. *Nature, Lond.*, **134**, 734.
COMPTON, A. H. and GETTING, I. A., 1935. *Phys. Rev.*, **47**, 817.

COMPTON, A. H. and TURNER, R. N., 1937. *Phys. Rev.*, **52**, 799.
COMPTON, A. H., WOLLAN, E. O. and BENNETT, R. D., 1934. *Rev. Sci. Instrum.*, **5**, 415.
CORLIN, A., 1930 a. *Lund. Obs.*, II, 58.
CORLIN, A., 1930 b. *Ark. Mat. Astr. Fys.*, **22**, B 2.
DRIGO, A., 1934. *Ric. Sci.*, **5**, 88.
DRIGO, A., 1935. *Ric. Sci.*, **6**, 529.
ECKART, C., 1934. *Phys. Rev.*, **45**, 851.
EHMERT, A., 1937. *Z. Phys.*, **106**, 751.
EULER, H. and WERGELAND, H., 1939. *Naturwissenschaften*, **27**, 484.
FERMI, E., 1940. *Phys. Rev.*, **57**, 485.
FORBUSH, S. E., 1937. *Phys. Rev.*, **51**, 1108.
FORBUSH, S. E., 1938. *Terr. Mag.*, **43**, 223.
FURRY, W. H., 1937. *Phys. Rev.*, **52**, 569.
FUSSELL, L., Jr., 1937. *Phys. Rev.*, **51**, 1005.
GEIGER, H. and ZEILLER, O., 1935. *Z. Phys.*, **97**, 300.
GILBERT, C. W., 1934. *Proc. Roy. Soc.*, A, **144**, 559.
GOCKEL, A., 1910. *Phys. Z.*, **11**, 280.
GOCKEL, A., 1911. *Phys., Z.*, **12**, 595.
GOLD, E., 1909. *Proc. Roy. Soc.*, A, **82**, 152.
GRANT, K., 1931. *Nature, Lond.*, **127**, 924.
GRAZIADEI, H., 1936. *Wiener Sitz. Ber.*, IIa, **145**, 495.
GREISEN, K., 1942. *Phys. Rev.*, **61**, 212.
GROSS, B., 1933. *Z. Phys.*, **83**, 214.
GUNN, R., 1933. *Terr. Mag.*, **38**, 247.
HAMILTON, J., HEITLER, W. and PENG, H. W., 1943. *Phys. Rev.*, **64**, 78.
HEISENBERG, W., 1939. *Z. Phys.*, **113**, 61.
HEITLER, W., 1936. *Quantum Theory of Radiation* (Oxford: University Press), p. 226.
HEITLER, W., 1938. *Proc. Roy. Soc.*, A, **166**, 529.
HERZOG, G., 1935. *J. Sci. Instrum.*, **12**, 153.
HESS, V. F, 1911. *Phys. Z.*, **12**, 998.
HESS, V. F., 1912 a. *Phys. Z.*, **13**, 1084.
HESS, V. F., 1912 b. *Wiener Sitz. Ber.*, IIa, **121**, 2001.
HESS, V. F., 1936. *Terr. Mag.*, **41**, 345.
HESS, V. F., and DEMMELMAIR, A., 1937. *Nature, Lond.*, **140**, 316.
HESS, V. F., DEMMELMAIR, A. and STEINMAURER, R., 1938. *Wiener Sitz. Ber.*, IIa, **147**, 89.
HESS, V. F., GRAZIADEI, H. and STEINMAURER, R., 1934. *Wiener Sitz. Ber.*, IIa, **143**, 313.
HESS, V. F., STEINMAURER, R. and DEMMELMAIR, A., 1938. *Nature, Lond.*, **141**, 686.
HOERLIN, H., 1936. *Z. Phys.*, **102**, 652.
HOFFMANN, G. and LINDHOLM, F., 1928. *Gerl. Beitr.*, **20**, 12.
HOFFMANN, G. and PFORTE, W. S., 1930. *Phys. Z.*, **31**, 347.
HUMMEL, J. N., 1934. *Naturwissenschaften*, **22**, 170.
JÁNOSSY, L., 1938. *Proc. Camb. Phil. Soc.*, **34**, 614.
JÁNOSSY, L. and ROCHESTER, G. D., 1941. *Nature, Lond.*, **148**, 531.
JÁNOSSY, L. and ROCHESTER, G. D., 1942. *Nature, Lond.*, **150**, 633.
JOHNSON, T. H., 1933. *J. Franklin Inst.*, **215**, 749.
JOHNSON, T. H., 1934. *Phys. Rev.*, **45**, 569.
JOHNSON, T. H., 1935. *Phys. Rev.*, **47**, 318.
JOHNSON, T. H., 1937. *J. Franklin Inst.*, **223**, 339.
JOHNSON, T. H., 1938 a. *Rev. Mod. Phys.*, **10**, 193.
JOHNSON, T. H., 1938 b. *Phys. Rev.*, **54**, 151.
JOHNSON, T. H., 1938 c. *Phys. Rev.*, **53**, 499.
JOHNSON, T. H., 1939. *Rev. Mod. Phys.*, **11**, 208.
JOHNSON, T. H. and BARRY, J. G., 1939. *Phys. Rev.*, **56**, 219.
JOHNSON, T. H., DE BENEDETTI, S. and SHUTT, R. P., 1943. *Rev. Sci. Instrum.*, **14**, 265.
JOHNSON, T. H. and POMERANTZ, M. A., 1939. *Phys. Rev.*, **55**, 104.
JOHNSON, T. H. and READ, D. N., 1937. *Phys. Rev.*, **51**, 557.
JOHNSON, T. H. and SHUTT, R. P., 1942. *Phys. Rev.*, **61**, 380.
JOHNSON, T. H. and STREET, J. C., 1933. *Phys. Rev.*, **43**, 381.
JONES, H., 1939. *Rev. Mod. Phys.*, **11**, 235.

KLEIN, O. and NISHINA, Y., 1928. *Nature, Lond.*, **122**, 399.
KOBAYASI, M. and OKAYAMA, T., 1939. *Proc. Phys.-Math. Soc., Japan*, **21**, 1.
KOLHÖRSTER, W., 1913. *Verh. dtsch. phys. Ges.*, **15**, 1111.
KOLHÖRSTER, W., 1914. *Verh. dtsch. phys. Ges.*, **16**, 719.
KORFF, S. A., 1939. *Rev. Mod. Phys.*, **11**, 211.
KORFF, S. A., 1941 a. *Rev. Sci. Instrum.*, **12**, 94.
KORFF, S. A., 1941 b. *Phys. Rev.*, **59**, 214.
KORFF, S. A., 1941 c. *Phys. Rev.*, **59**, 949.
KORFF, S. A., 1941 d. *Proc. Amer. Phil. Soc.*, **84**, 589.
KORFF, S. A. and CLARKE, E. T., 1942. *Phys. Rev.*, **61**, 422.
LEMAÎTRE, G. and VALLARTA, M. S., 1933. *Phys. Rev.*, **43**, 87.
LOCHER, G. L., 1936. *Rev. Sci. Instrum.*, **7**, 471.
MAIER-LEIBNITZ, H., 1939. *Z. Phys.*, **112**, 569.
MCLENNAN, J. C. and BURTON, E. F., 1903. *Phys. Rev.*, **16**, 184.
MILLIKAN, R. A. and BOWEN, I. S., 1923. *Phys. Rev.*, **22**, 198.
MILLIKAN, R. A. and BOWEN, I. S., 1926. *Phys. Rev.*, **27**, 353.
MILLIKAN, R. A. and CAMERON, G. H., 1926. *Phys. Rev.*, **28**, 851.
MILLIKAN, R. A. and CAMERON, G. H., 1928. *Phys. Rev.*, **31**, 163.
MILLIKAN, R. A. and CAMERON, G. H., 1931. *Phys. Rev.*, **37**, 235.
MILLIKAN, R. A. and NEHER, H. V., 1936. *Phys. Rev.*, **50**, 15.
MILLIKAN, R. A., NEHER, H. V. and PICKERING, W. H., 1942. *Phys. Rev.*, **61**, 397.
MILLIKAN, R. A. and OTIS, R. M., 1926. *Phys. Rev.*, **27**, 645.
MØLLER, C., 1932. *Ann. Phys., Lpz.*, **14**, 531.
MONTGOMERY, C. G. and MONTGOMERY, D. D., 1935. *Phys. Rev.*, **47**, 429.
MONTGOMERY, C. G. and MONTGOMERY, D. D., 1936. *Phys. Rev.*, **49**, 705.
MONTGOMERY, C. G. and MONTGOMERY, D. D., 1939. *Phys. Rev.*, **56**, 10.
MONTGOMERY, C. G. and MONTGOMERY, D. D., 1941 a. *J. Franklin Inst.*, **231**, 447.
MONTGOMERY, C. G. and MONTGOMERY, D. D., 1941 b. *J. Franklin Inst.*, **231**, 509.
MONTGOMERY, C. G., MONTGOMERY, D. D., RAMSEY, W. E. and SWANN, W. F. G., 1936. *Phys. Rev.*, **50**, 403.
MONTGOMERY, C. G., RAMSEY, W. E., COWIE, D. B. and MONTGOMERY, D. D., 1939. *Phys. Rev.*, **55**, 1117.
MORGAN, K. Z. and NIELSEN, W. M., 1937. *Phys. Rev.*, **52**, 564.
MOTT-SMITH, L. M., 1934. *Rev. Sci. Instrum.*, **5**, 346.
NEDDERMEYER, S. H. and ANDERSON, C. D., 1937. *Phys. Rev.*, **51**, 884.
NEDDERMEYER, S. H. and ANDERSON, C. D., 1938. *Phys. Rev.*, **54**, 88.
NEDDERMEYER, S. H. and ANDERSON, C. D., 1939. *Rev. Mod. Phys.*, **11**, 191.
NEHER, H. V. and PICKERING, W. H., 1938. *Phys. Rev.*, **53**, 111.
NIELSEN, W. M., 1938. *J. Franklin Inst.*, **226**, 601.
NIELSEN, W. M., MORGAN, J. E. and MORGAN, K. Z., 1939. *Phys. Rev.*, **55**, 995.
NISHINA, Y., TAKEUCHI, M. and ICHIMIYA, T., 1939. *Phys. Rev.*, **55**, 585.
NORDHEIM, L. W., 1936. *Phys. Rev.*, **49**, 189.
NORDHEIM, L. W. and HEBB, M. H., 1939. *Phys. Rev.*, **55**, 1111.
NORDHEIM, L. W. and NORDHEIM, G., 1938. *Phys. Rev.*, **54**, 254.
OPPENHEIMER, J. R., 1935. *Phys. Rev.*, **47**, 44.
PEIERLS, R., 1939. *Rep. Prog. Phys.*, **6**, 78.
PFOTZER, G., 1936. *Z. Phys.*, **102**, 23.
PRIEBSCH, J. A., 1935. *Z. Phys.*, **95**, 102.
PRIEBSCH, J. A., 1936. *Wien. Ber.*, **145**, 101.
RAMSEY, W. E., 1942. *Phys. Rev.*, **61**, 96.
RAMSEY, W. E. and HUDSPETH, E. L., 1942. *Phys. Rev.*, **61**, 95.
RASETTI, F., 1941. *Phys. Rev.*, **59**, 706.
REGENER, E., 1932. *Naturwissenschaften*, **20**, 695.
REGENER, E. and PFOTZER, G., 1935. *Nature, Lond.*, **136**, 718.
ROCHESTER, G. D., 1944. *Nature, Lond.*, **154**, 399.
ROSE, M. E. and KORFF, S. A., 1941. *Phys. Rev.*, **59**, 850.
ROSE, M. E. and RAMSEY, W. E., 1942. *Phys. Rev.*, **61**, 198.
ROSSI, B., 1930 a. *Nature, Lond.*, **125**, 636.

Rossi, B., 1930 b. *Phys. Rev.*, **36**, 606.
Rossi, B., 1931 a. *Nuovo Cimento*, **8**, 85.
Rossi, B., 1931 b. *Rend. Lencei*, **13**, 47.
Rossi, B., 1933. *Z. Phys.*, **82**, 151.
Rossi, B., 1939. *Rev. Mod. Phys.*, **11**, 296.
Rossi, B., 1940. *Phys. Rev.*, **57**, 469.
Rossi, B. and De Benedetti, S., 1934. *Ric. Sci.*, **5**, 379.
Rossi, B. and Greisen, K., 1942. *Phys. Rev.*, **61**, 121.
Rossi, B., Hilberry, N. and Hoag, J. B., 1940. *Phys. Rev.*, **57**, 461
Rossi, B. and Regener, V. H., 1940. *Phys. Rev.*, **58**, 837.
Rutherford, E. and Cooke, H. L., 1903. *Phys. Rev.*, **16**, 183.
Sabato, E. R., 1939. *Phys. Rev.*, **55**, 1272.
Schein, M., 1944. *Phys. Rev.*, **65**, 344.
Schein, M. and Gill, P. S., 1939. *Rev. Mod. Phys.*, **11**, 267.
Schein, M., Jesse, W. P. and Wollan, E. O., 1940. *Phys. Rev.*, **57**, 847.
Schein, M., Jesse, W. P. and Wollan, E. O., 1941. *Phys. Rev.*, **59**, 615.
Schein, M. and Wilson, V. C., 1939 a. *Rev. Mod. Phys.*, **11**, 292.
Schein, M. and Wilson, V. C., 1939 b. *Phys. Rev.*, **54**, 304.
Schindler, H., 1931. *Z. Phys.*, **72**, 625.
Schmeisser, K., 1938. *Z. Phys.*, **110**, 443.
Schmeisser, K., 1939. *Z. Phys.*, **112**, 501.
Schmeisser, K. and Bothe, W., 1938. *Ann. Phys., Lpz.*, **32**, 161.
Schonland, B. F. J., Delatinzky, B. and Gaskell, J., 1937. *Terr. Mag.*, **42**, 137
Schwegler, A., 1936. *Z. Phys.*, **101**, 93.
Shutt, R. P., De Benedetti, S. and Johnson, T. H., 1942. *Phys. Rev.*, **62**, 552.
Steinke, E., 1927. *Z. Phys.*, **42**, 570.
Steinke, E., 1928. *Z. Phys.*, **48**, 647.
Stoermer, C., 1904. *Vid. Selsk. Skr., Christiania.*
Stoermer, C., 1913. *Vid. Selsk. Skr., Christiania.*
Stoermer, C., 1930. *Z. Astrophys.*, **1**, 237.
Stoermer, C., 1931. *Ergeb. d. Kosmischen Phys.*, **1**, 1.
Stoermer, C., 1932. *Terr. Mag.*, **37**, 375.
Stoermer, C., 1934. *Univ. Obser. Oslo* (1), Pub. No. 10, 5.
Street, J. C., 1939. *J. Franklin Inst.*, **227**, 765.
Street, J. C. and Stevenson, E. C., 1936. *Rev. Sci. Instrum.*, **7**, 347.
Street, J. C. and Stevenson, E. C., 1937 a. *Phys. Rev.*, **52**, 1003.
Street, J. C. and Stevenson, E. C., 1937 b. *Phys. Rev.*, **51**, 1005.
Street, J. C. and Young, R. T., Jr., 1934. *Phys. Rev.*, **46**, 823.
Swann, W. F. G., 1924. *Phil. Mag.*, **47**, 306.
Swann, W. F. G., 1925. *J. Franklin Inst.*, **200**, 489.
Swann, W. F. G., 1927. *Phil. Mag.*, **3**, 1088.
Swann, W. F. G., 1930. *J. Franklin Inst.*, **209**, 151.
Swann, W. F. G., 1932. *Phys. Rev.*, **42**, 914.
Swann, W. F. G., 1933 a. *Phys. Rev.*, **44**, 224.
Swann, W. F. G., 1933 b. *Phys. Rev.*, **44**, 124.
Swann, W. F. G., 1933 c. *Phys. Rev.*, **43**, 217.
Swann, W. F. G., 1933 d. *J. Franklin Inst.*, **215**, 273.
Swann, W. F. G., 1934 a. *Phys. Rev.*, **46**, 828.
Swann, W. F. G., 1934 b. *Phys. Rev.*, **45**, 295.
Swann, W. F. G., 1935. *J. Franklin Inst.*, **220**, 373.
Swann, W. F. G., 1936 a. *Phys. Rev.*, **50**, 1103.
Swann, W. F. G., 1936 b. *Phys. Rev.*, **49**, 478.
Swann, W. F. G., 1937. *Phys. Rev.*, **51**, 718.
Swann, W. F. G., 1938 a. *J. Franklin Inst.*, **226**, 757.
Swann, W. F. G., 1938 b. *J. Franklin Inst.*, **226**, 598.
Swann, W. F. G., 1938 c. *J. Franklin Inst.*, **226**, 441.
Swann, W. F. G., 1939 a. *Rev. Mod. Phys.*, **11**, 242.
Swann, W. F. G., 1939 b. *Phys. Rev.*, **56**, 209.

SWANN, W. F. G., 1940. *Phys. Rev.*, **58**, 200.
SWANN, W. F. G., 1941. *Phys. Rev.*, **59**, 770.
SWANN, W. F. G., 1943. *J. Franklin Inst.*, **236**, 1.
SWANN, W. F. G. and DANFORTH, W. E., 1939. *J. Franklin Inst.*, **228**, 43.
SWANN, W. F. G. and LOCHER, G. L., 1935. *Nat. Geog. Soc. Cont. Tech. Papers, Stratosphere Series*, No. 1, 7.
SWANN, W. F. G., LOCHER, G. L. and DANFORTH, W. E., 1936 a. *Nat. Geog. Soc. Cont. Tech. Papers, Stratosphere Series*, No. 2, 13.
SWANN, W. F. G., LOCHER, G. L. and DANFORTH, W. E., 1936 b. *J. Franklin Inst.*, **221**, 275.
SWANN, W. F. G., LOCHER, G. L. and DANFORTH, W. E., 1937. *Phys. Rev.*, **51**, 389.
SWANN, W. F. G. and MONTGOMERY, C. G., 1933 a. *Phys. Rev.*, **43**, 782.
SWANN, W. F. G. and MONTGOMERY, C. G., 1933 b. *Phys. Rev.*, **44**, 52.
SWANN, W. F. G. and RAMSEY, W. E., 1938. *Phys. Rev.*, **54**, 229.
SWANN, W. F. G. and RAMSEY, W. E., 1939. *Phys. Rev.*, **56**, 378.
SWANN, W. F. G. and RAMSEY, W. E., 1940 a. *Phys. Rev.*, **57**, 1051.
SWANN, W. F. G. and RAMSEY, W. E., 1940 b. *Phys. Rev.*, **57**, 749.
SWANN, W. F. G. and RAMSEY, W. E., 1940 c. *Phys. Rev.*, **58**, 477.
THOMPSON, J. L., 1938. *Phys. Rev.*, **55**, 11.
THOMSON, J. J., 1912. *Phil. Mag.*, **23**, 449.
TUVE, M. A., 1930. *Phys. Rev.*, **35**, 651.
VOS, P. J. G. DE, 1940. *Nature, Lond.*, **145**, 387.
WEISZ, P. B., 1941. *Phys. Rev.*, **59**, 845.
WEISZ, P. B., 1942. *Phys. Rev.*, **62**, 477.
WEISZ, P. B. and RAMSEY, W. E., 1942. *Rev. Sci. Instrum.*, **13**, 258.
WILLIAMS, E. J., 1932. *Proc. Roy. Soc.*, A, **135**, 108.
WILLIAMS, E. J. and ROBERTS, G. E., 1940. *Nature, Lond.*, **145**, 102.
WILSON, V. C., 1938 a. *Phys. Rev.*, **53**, 337.
WILSON, V. C., 1938 b. *Phys. Rev.*, **53**, 908.
WOODWARD, R. H., 1936. *Phys. Rev.*, **49**, 711.
WOODWARD, R. H. and STREET, J. C., 1935. *Phys. Rev.*, **47**, 800.
WOODWARD, R. H. and STREET, J. C., 1936. *Phys. Rev.*, **49**, 198.
YUKAWA, H., 1935. *Proc. Phys.-Math. Soc., Japan*, **17**, 48.
YUKAWA, H., SAKATA, S., KOBAYASI, M. and TAKETANI, M., 1938. *Proc. Phys.-Math. Soc., Japan*, **20**, 720.
ZANSTRA, H., 1934. *Naturwissenschaften*, **22**, 171.

SEISMOLOGY

By HAROLD JEFFREYS, F.R.S.,
St. John's College, Cambridge

§1. INTRODUCTION

The modern study of seismology falls into three main divisions. The first concerns the local effects of earthquakes, especially their destructive effect, which connects the subject with engineering, and the observed displacements of the Earth's surface, which connect it with structural geology. As a scientific study it is almost as old as astronomy, instruments for estimating the direction and intensity of the movement having been known to the ancient Chinese. The second division concerns the effects of earthquakes outside the region of destruction. These take the form of elastic waves, some of which travel through the interior of the earth and some over its surface, and are recorded by suitable instruments. These waves give us the most detailed information about the internal structure of the Earth that we have. The third concerns artificial earthquakes. The methods devised for the analysis of the motion due to small earthquakes, accurately recorded at short distances, are applicable to the waves sent out when a buried charge of high explosive is released. They have been used to determine the structure of the upper layers, mainly sedimentary, and especially to locate buried anticlines in oil-bearing rocks and hence to indicate the best places to make borings.

Popular interest in seismology concerns only the first division, and even there its distribution is strangely erratic. The occasional reports of earthquakes in the Press give an impression that earthquakes are much less frequent than they actually are. An indication of their actual frequency is given by the *International Seismological Summary*, which collects the information from most of the existing observatories. The condition for including an earthquake in the *Summary* is roughly that it should be recorded by the observatories in more than one country, and on an average about 600 are included per year. This excludes earthquakes, specially prevalent in Japan and California, that are too small to be recorded except locally; these may reach thousands per year.

Modern developments are almost entirely concerned with the applications of the theory of elastic waves. The prediction that an elastic solid should be able to transmit two types of wave through its substance was first made by Poisson in 1829. One of these is similar to the sound wave in a liquid or gas; when it passes a particle, that particle vibrates in the direction of travel of the wave. Such a wave is called longitudinal or compressional, the latter term being used because the passage of the wave is associated with changes of volume. The second type can occur only in a solid; when it passes a particle, the particle is displaced at right angles to the direction of travel of the wave. It is often called transverse or distortional. The usual practice is to denote the longitudinal waves by P (for *primary*, because they have the greater speed) and the transverse ones by S, for *secondary*. Turner suggested that P and S might be taken to

indicate *push* and *shake*, though P is as often as not a *pull*. P and S can travel in any direction. Their velocities α and β are given, in terms of the elastic properties and density of the material, by

$$\alpha^2=(\lambda+2\mu)/\rho\ ;\quad \beta^2=\mu/\rho,$$

where ρ is the density, μ the rigidity, and $\lambda+\frac{2}{3}\mu$ the bulk-modulus or the reciprocal of the compressibility. For the type of solid considered by Poisson, $\lambda=\mu$ and $\alpha=\beta\sqrt{3}$. These relations are roughly correct for rocks.

It was first recognized by Rayleigh that another type of wave can travel over the surface with a velocity rather less than β, the amplitude diminishing with depth and becoming inappreciable at depths more than a few wave-lengths, as for water waves on deep water. Another type of surface wave, found by observation, was explained by Love in 1911.* It is a transverse wave that can exist if the velocity of S increases with depth. In spite of their apparent similarity, surface waves and water waves are theoretically quite different. Water waves are controlled by gravity. The surface waves of earthquakes are elastic waves, and gravity has little or no effect on them.

The theory of the propagation of elastic waves is similar to that of other wave motions, the following properties being specially important. (i) S shows polarization, as for light waves ; displacements in two directions in the wave front and perpendicular to each other are propagated independently, with the same velocity. It is usual to take one of these directions to be horizontal and the other to be in the plane containing the ray and the centre of the Earth, and to denote the components by SH and SV. (ii) Times of travel, rays and wave fronts are related according to the principle of stationary time, and in general the distribution of energy over the wave front is approximately the same as if the energy within a given cone of rays travelled without influence from neighbouring cones. (iii) When an SH wave meets a free surface it is wholly reflected as SH ; when one meets a horizontal interface it follows the usual law of refraction, but there is also a reflected wave, reflexion becoming total at a critical angle. The behaviour of P and SV is more complicated because each requires two boundary conditions at a free surface and four at an interface, instead of one and two respectively as in the cases of sound and light. When either meets a free surface, reflected waves of both types are produced ; when either meets a horizontal interface it produces refracted waves and reflected waves of both types. There are critical angles in both cases, and some of the derived movements, when the angle of incidence is high, are not waves travelling away from the interface. (iv) If a transition between two media is spread uniformly, or nearly so, through a region several wave-lengths in extent, waves reaching it are almost completely transmitted in the original type, or, if this is forbidden by the law of refraction, they are totally reflected as in atmospheric mirage. (v) The wave velocities of bodily waves are practically independent of the wave-length. This is not true of surface waves, and considerable dispersion occurs.

A disturbance that is zero up to a definite instant is called a *pulse* or a *phase*.

A wave may be reflected several times at the outer surface. The history of each is indicated by the notation. Thus PP is a wave that starts as P and is

* This, and other earlier references, appear in *The Earth* (Jeffreys, 1929).

once reflected as P on the way to the observatory. PS is one that starts as P and is reflected as S. PPP has been twice reflected, remaining of P type throughout, and similarly for other phases.

The systematic recording of earthquake waves began in the later nineteenth century, especially in Japan and Italy. The earlier instruments were not very satisfactory. The ideal would be to record the displacement of the ground in each direction against the time, but this is mechanically impossible. A fair approximation to it over a range of periods is possible, and has been achieved in many ways. The most important property in a recording instrument is a quick response to any sudden change of the velocity of the ground, so that times of arrival of the different phases can be read accurately. The next is a quick recovery after a displacement, so that the instrument will be ready to record later movements. These require the instrument to have a suitable natural period and damping. It is best for the natural period to be comparable with that of the first swings ; about 7^s is convenient, and damping should be enough to make the second swing after an impulse small compared with the first. Shorter periods, such as 1^s or 2^s, are found useful for some purposes. Some of the earlier instruments had little damping and small magnification, and it is probably for this reason that it was so long before the theoretical predictions of Poisson and Rayleigh were verified ; P, S and the surface waves were first clearly distinguished, in the records of an Indian earthquake, by R. D. Oldham in 1900. The times of P and S did not increase with distance as fast as they would if the velocities were the same at all depths. This implies that the velocities increase with depth, which implies in turn that the rays are curved upwards. But when the times of travel between points on the surface of a sphere are known, it is a soluble problem to find what distribution of velocity with depth will give them. This problem was solved formally by Herglotz, Bateman and Wiechert. It remained to find the travel times ; this could be done only by collection and analysis of large numbers of observations. Oldham made a beginning, but his tables were superseded by those of K. Zöppritz (1907). Many other workers attempted to improve on Zöppritz's tables, notably A. and S. Mohorovičić and B. Gutenberg. Substantial differences remained outstanding, and it is only within the last few years that it has been possible to say that the travel times of P and S are known within the ordinary inaccuracy of reading.

The place where the earthquake actually occurs is called the *focus*, and the point on the surface vertically above it is the *epicentre*. The opposite point of the surface is called the *anticentre*. The distance from the epicentre to an observing station is usually calculated as an angle. Until recently it was sufficiently accurate for seismological purposes to treat the Earth as a sphere, and the distance was taken as the angle between the verticals at the epicentre and the station. It has now been found better to take it as the angle Δ subtended at the centre of the Earth by the line joining the epicentre and the station ; for a spherical Earth the two would be identical, but modern discussions have reached such an accuracy that the difference must be taken into account. The need for the distinction was pointed out independently by Gutenberg and Richter and by Comrie. The two distances are known respectively as the *geographic* and *geocentric* distances.

§2. THE CORE

Oldham noticed in 1906 that, while P reaches a distance of 100° in about 14^m, it takes about 20^m to reach 180°. He found that the difference was too large to be attributed to distance alone; it was necessary that there should be a drop of velocity at the greater depths. The consequences of this were worked out, partly by Oldham and very fully by Gutenberg (1914). The velocity of P at short distances was about 8 km./sec., rising to about 13 km./sec. at a depth of about 0·4 of the radius, but to fit the time at 180° it must then drop to about 9 km./sec. The result would be that a ray reaching this depth would be sharply refracted downwards, and there would be a shadow zone. The nearer edge of the shadow was located by Gutenberg, by study of individual seismograms, at about 103°, and the further edge at about 143° where P reappeared with an enormous amplitude, falling off rapidly at greater distances. The phenomenon can be imitated optically by filling a spherical flask with water and placing a pinhole source of light and a photographic plate on opposite sides of it at a distance of about twice the radius. The ratio of the velocities of light in air and water is about the same as that of the velocities of P just outside and within the core. The rays of light are cut off by the plate before they reach the focus, but there is a caustic surface surrounded by a shadow. The plate shows, therefore, a very bright ring of light, corresponding to the large intensity about 143°; outside this is shadow; inside it there is illumination, but less intense.

The P wave that has passed through the central core is now usually denoted by PKP, K by itself being used for the phase inside the core. Oldham's discovery by itself would not determine whether the boundary of the core is a discontinuity or a continuous transition, but it has been shown by other methods that the former is correct. Gutenberg proceeded to an extensive calculation of the times of other phases that should exist if the boundary is a discontinuity. A P or SV wave striking an internal discontinuity would give rise to two reflected and two refracted waves, one P and one SV of each. The reflected P and S waves are known as PcP and ScS (the c being needed to distinguish them from the surface reflexions PP and SS). But an S wave striking the core should give rise to P and S waves in the core, if the core is solid. These would be broken up afresh on striking the core boundary again, giving a great variety of phases, which should reach the surface at different times and distances. The most important of these is SKS, which is of S type in both passages through the outer shell, but of P type in the core. Gutenberg identified this phase and several other core phases; but all that he found had been of P type in the core. He also tried to find waves that had been of S type in the core, assuming the ratio of velocities to be about the same as in the shell (about 1·8), but failed. Several other workers, notably Macelwane (1930) and Bastings (1935), have claimed to detect such waves, but the results are conflicting. Some observations seem to refer to SKKS, some to SKSP, some to PPS, and some to the rather complicated wave ScSPKP. (The last, as Gutenberg has pointed out, has a caustic like PKP, and, therefore, is specially liable to be mistaken for S through the core. Its existence was first pointed out, and approximate times were first given for it, by Gutenberg and Richter (1934, p. 123).) Others are doubtfully readings of genuine phases

at all. There is no evidence that the core can transmit S waves, and if it cannot it is liquid. We shall see later that there is other and more direct evidence in favour of this suggestion.

At first the most prominent phase that undergoes reflexion at the core boundary was SKKS (reflected, of course, on the inside). Its existence provided good evidence that the core boundary was a discontinuity. The simplest reflexions, PcP and ScS, were not clearly recognized for some time. This was probably because they were mixed up with the train of surface waves ; at any rate the first published readings of them seem to have been in a paper on a deep-focus earthquake by F. J. Scrase (1933), in which surface waves were small, and Gutenberg and Richter have recently found them readable in normal earthquakes on the records made by short-period instruments. Their existence is decisive evidence that the core boundary is a discontinuity.

A possibility that does not seem to have been discussed is that PKP might really be of S type in the core and that the true P wave in the core is faster, so that the search should really have been for waves that travel in the core at about 15 km./sec. This, however, is impossible for two reasons. A wave at normal incidence would give zero amplitude for the wave of the other type. Hence PKP at the anticentre, having undergone two such transformations, would have zero amplitude. This is not so. Also, on this hypothesis, a wave that has been of P type all the way would still be expected at the anticentre, but much earlier than PKP. This is not observed. Gutenberg's assumption that PKP is of P type all through is, therefore, the only possible one.

§3. NEAR EARTHQUAKES

At distances up to 700 km. the records of an earthquake are strikingly different from those at larger distances. The oscillations are much more rapid, a single swing and recovery taking about 1^s or 2^s as against 4^s to 7^s at larger distances. To get useful information it is necessary to study rather small earthquakes, since the disturbances at short distances in large earthquakes are too violent for accurate reading. It was discovered in 1909 by A. Mohorovičić that the records of an earthquake in Croatia showed two P and two S phases, travelling with different velocities. He inferred that there was an upper layer in the crust where the velocities were substantially less than below it. A fracture within this layer would send out P and S waves ; these would in part travel directly to the surface, and in part would be refracted into the lower layer, travel along in this, and be refracted up again when they again strike the interface. This might have been expected from geological considerations, since most igneous rocks fall into two main types, according as they contain free silica or not, and the former appear to have come from the smaller depths. The direct waves are denoted by Pg and Sg, P and S by themselves being reserved for waves that have been in the lower layer. At very short distances (say up to 150 km.) P arrives after Pg and S after Sg, which are considerably larger, and P and S are unreadable ; but at greater distances they precede Pg and Sg respectively and can be read. At these short distances the travel times are practically linear functions of the distance, but the constant terms differ on account of the time lost by P and S in the inclined parts of their paths, and lead to information relevant to the thickness

of the upper layer. Gutenberg (1915) made a very thorough examination of the records of two earthquakes in South Germany, which occurred in 1911 and 1913, and obtained similar results from a much larger amount of observational material. V. Conrad (1923) found another traceable P pulse, which appeared to have travelled in an intermediate layer ; this was also found by the present writer (1927) in the records of the Jersey earthquake of 1926, and the corresponding S pulse was also found. This pair are denoted by P* and S*.

Unfortunately most of the analyses for near earthquakes are completely vitiated by inadequate statistical treatment. It is obvious that if the focus is at a depth h and the horizontal distance is x, the time of Pg would be $(x^2+h^2)^{\frac{1}{2}}/c$, where c is the velocity, assumed uniform. If the velocity varies with depth, the formula is more complicated. But inspection of the data shows that a linear law fits the data very well, within departures of the order of a second. To establish the possibility of finding h or the variation of velocity with depth from the data it would be necessary to make a least-squares solution and show that the estimates of these quantities considerably exceed their standard errors. In practice continental workers have proceeded by a number of different methods, largely graphical, and given estimates of about 35 km. for the focal depth and 55 km. for the depth down to the top of the lower layer, usually with stated uncertainties of a few kilometres. These continue to be quoted in text-books even though their baselessness has been repeatedly pointed out. This is not the place to give the objections in detail, but it may be mentioned that if the times of Pg and P had been calculated from the solutions given, to give a proper final check, it would have been found that the whole of the calculated P times were about 4^s later than the observed times that they are alleged to be based on. In one case I found a complete disagreement between the observations of a pulse in an intermediate layer and the solution given ; it turned out afterwards that the author had obtained his solution graphically and miscounted the large squares. But that solution has been republished in a text-book without reference to my corrected solution. Perhaps the phenomenon of good observations being maltreated in the process of reduction so as to give a completely worthless result is not unknown in other branches of physics.

Least-square determinations of the velocities of Pg, Sg, P and S in numerous near earthquakes are in satisfactory agreement. The velocities in the intermediate layer are much more difficult to determine. Conrad's observations of P* in the Schwadorf earthquake agreed with mine in the Jersey earthquake. His earlier results for the Tauern earthquake gave a significantly smaller velocity, but some of the readings are capable of other interpretations (Jeffreys, 1937 b).

Gutenberg (1932) finds two different intermediate layer P waves in California, neither agreeing with any European determination. Byerly and Wilson (1935), however, working in the same region, get different results again. The problem here is one of difficulty of reading. Pg and Sg are large sharp movements, and are habitually read at the stations themselves in their routine measurement. But between P and Pg, and between S and Sg, the record shows a continual oscillation of fluctuating amplitude. It is difficult to be sure which of the increases in amplitude are sufficiently sudden to be regarded as definite phases, and there is a possibility that personal opinion may have played a part. On the other hand, in many Japanese earthquakes long series of routine readings, sometimes at as

many as twenty stations, are in good agreement with the times of S* calculated from the Jersey earthquake velocity. There seem to be systematic differences between regions, but these Japanese data seem to show that the readings for intermediate-layer waves correspond to genuine phases and are not accidental.

Determinations of the focal depth and the thicknesses of the upper and intermediate layers from near earthquakes are difficult. Focal depth could be shown only by an upward curvature at short distances in the time-curve for Pg; the departure from a straight line would hardly be appreciable except at the nearest station, and is comparable with the uncertainty of observation. Schmerwitz (1938 ; cf. also Jeffreys, 1939 d) has made a statistical study of a number of recent earthquakes in the Alps and South Germany, and finds in several cases that the estimate of h^2 would be negative. This is just what would be expected if h was too small to be determined from the data ; if we try in such conditions to determine it, it is as likely as not that a least-squares solution for h^2 will give a negative estimate. All that we can say is that if Pg exists the focus is certainly in the upper layer, and if linear forms are fitted to the times of Pg and P*, or Sg and S*, the constant terms will differ by amounts proportional to the vertical distance travelled by P* or S* in the upper layer. Thus we get extreme cases according as the focus is supposed to be at the surface or at the base of the upper layer ; the thickness calculated in the latter case will be twice that in the former, and the truth must lie between them. From such considerations it appears that the thickness of the upper layer in Europe is between 19 and 33 km. That of the intermediate layer is found to be 9 ± 3 km. (Jeffreys, 1937 b, p. 210). But there are some outstanding difficulties in the interpretation of the data ; the constant terms in the times of the various pulses do not agree with any hypothesis as closely as the consistency of a single series of observations would lead us to expect.

§ 4. SURFACE WAVES

The dispersion of Love and Rayleigh waves provides useful additional information about the structure of the upper layers. On account of the fact that they do not spread downwards, they lose less amplitude with distance than the bodily waves do, and at large distances they form the most prominent feature of the record. Their prevailing periods are 10^s to 30^s. On a uniform solid the velocity of Rayleigh waves would be the same for all periods, and Love waves would not exist at all. The two types can be distinguished by the fact that in Rayleigh waves the displacement of a particle is partly vertical and partly in the plane of travel ; in Love waves the displacement is at right angles to the plane of travel. The distinction is similar to that between P and SV on the one hand and SH on the other. Consequently if we had all three components of the motion the two types could be separated completely by a rather laborious analysis. But this can be avoided because the vertical component shows only Rayleigh waves, and if we choose a station such that the waves come to it from nearly north or nearly west, one horizontal component will show the Rayleigh waves and the other the Love waves.

The seismograms actually show on each component a long train of approximately simple harmonic waves, varying slowly in amplitude and period. In other words the surface waves show dispersion. This was explained by Love

as due to the heterogeneity of the crust. If the velocity of S increases with depth, Love found that for short waves of both types the wave velocity would be determined mainly by that of S at small depths, and conversely. The existence of dispersion in the surface waves is therefore evidence for a layered crust, confirming the near earthquake evidence. The conditions of observation make it impossible to follow an individual wave, but we can find the rate of travel of each particular period. This is the group-velocity associated with that period, and if we work out the variation of group-velocity with period according to various possible structures and compare the results with observation we can considerably limit the range of structure permitted. Most of the work on these lines has been done by Stoneley (1928, 35, 37). The observational difficulty is much less than for near earthquake studies, but the labour of computation is far greater. This method will probably be considerably extended in future. So far it has been usual to assume a particular ratio of the thicknesses of the upper and intermediate layers and get an absolute determination by matching the group-velocities for a particular period ; but greater detail would be possible.

So far the surface waves provide our only method of investigating the ocean floor at small depths. They have the great advantage that they can travel for any distance and do not require a large number of stations near the epicentre. The maximum velocity appears to be about the same as in the continents, but a given fraction of it occurs at a smaller period, indicating that the upper layer is thinner (Byerly, 1930 ; Bullen, 1939). Little is known at present about the velocities under the ocean for short periods, and such information would be of great value.

§5. THE INTERNATIONAL SEISMOLOGICAL SUMMARY

There are approximately 500 seismological observatories in the world. Many of these are purely seismological, but others carry on seismology in association with astronomy or meteorology. The observer at each station reads the most prominent phases on his record of each earthquake. Little use, however, can be made of a single record in comparison with what can be got from the comparison of a large number. When the announcement of a large earthquake appears in the newspapers, with a statement of the position of the epicentre given by a single station, it has usually been obtained as follows. The interval between P and S increases with distance, and hence the observed interval gives an estimate of the distance of the epicentre from the station. The ratio of the extents of the first P movements on the horizontal components determines the direction, with an ambiguity of sign. This cannot be resolved without further information, since the motion may be either inwards or outwards from the epicentre, and the observer sometimes takes the more seismic region as the most likely. Information from one more station will settle the matter, and so will the vertical component if it is recorded ; for if the displacement is outwards the vertical displacement will be upwards, and conversely. Such determinations are only preliminary and cannot be accurate because the displacement is difficult to measure. Errors of 10° are quite normal. Accurate determinations need the comparison of widely separated stations.

The systematic collection of the world's information was begun by John Milne and carried on by H. H. Turner at Oxford, under the auspices of the British Association. In 1922 the scheme was approved by the International Seismological Association, which contributed a large share of the cost. Up to Turner's death in 1930 he was assisted by J. S. Hughes and Miss E. F. Bellamy, who have continued the work under the supervision of Prof. H. H. Plaskett. Most stations send their readings to Oxford, and when the data for an earthquake have all arrived, its position and time are worked out. The results, including a comparison of each observed time with the calculated time, are published quarterly in the International Seismological Summary.* The solutions were originally based on the intervals between P and S at the stations, but it has now been found better to use the times of P alone. No attempt is made to get the best possible solution from the whole of the data, since it would take too long; a solution is made from the time of P at a few stations, and the residuals at all stations are available for a more accurate determination if one is required.

§ 6. RECENT WORK ON TRANSMISSION TIMES

The Zöppritz tables of P and S, with supplementary tables computed by Turner, were used for solution and comparison in this work for earthquakes up to the end of 1929 (published in 1933). Turner, however, published an important paper in 1926, in which he classified the residuals for 1918–22 by distance, and found that they varied systematically. The means varied with distance by about 20s for P and 30s for S. About the same time P. Byerly (1926) gave a set of times based on his special study of the Montana earthquake of 1925 June 28. I compared these with the corrections found by Turner and found very close agreement (Jeffreys, 1928). It appeared at this stage that these corrections were not likely to be far wrong. In two later papers (1931 b) I extended Turner's work to the earthquakes of 1923 to March 1927, and again got similar corrections, which are embodied in my 1932 tables, published by the British Association Seismological Committee. The apparent standard errors of these times were mostly under a second, but there were grounds for mistrusting them. The corrections were simple means of the residuals (apart from a correction to take account of departure from the normal law of error) against the Zöppritz times; but the epicentres and times of occurrence had been found so as to give the best possible agreement with the latter, and in different earthquakes the observations were differently distributed in distance. Thus part of the variation of the errors in the tables would have been converted into errors of the epicentre and times of occurrence, and would not be revealed by a mere classification of the residuals. I called attention to this feature at the time (1931 b, p. 337), but some critics overlooked the warning. The Zöppritz times being taken as a first approximation, my 1932 ones were a second approximation, but it would be necessary to complete the process before the apparent accuracy could be claimed as genuine.

At this stage I obtained the valuable assistance of K. E. Bullen. It was necessary (Jeffreys and Bullen, 1935) to reduce afresh all the earthquakes used

* Referred to hereafter as the I.S.S.

for P and S, determining corrected epicentres and times of occurrence so as to give the best possible fit with the 1932 tables, and then to classify the new residuals. This showed that further corrections, reaching about 5ˢ for P and 9ˢ for S, were still needed. Applying these, we got a new set of tables and repeated the process till there was no further change. S was not found useful in determining epicentres at this stage, and the solutions for the individual earthquakes rest on P alone. Other phases were not treated until the solution for P was complete.

For S there were several troubles. Up to about 20°, the S residuals were spread over about 20ˢ without any convincing concentration of frequency, and our final solution was based very largely on analogy with the times of P in this range, combined with knowledge of the velocity of S at short distances derived from special studies by Conrad and myself. This scatter of the S residuals remains unexplained. From 25° to 83° it was found that the S residuals for a given earthquake varied fairly smoothly, but in comparison of different earthquakes there was a complication, which we called the Z phenomenon. S behaved as if its time of origin differed from that of P, and not always by the same amount. The time of origin found from P being t_0, the average S residual, taking the same t_0, might be anything from -12^s to $+8^s$. This was noticed first by Lehmann and Plett (1932) in their studies of the Peru and Marianne Islands earthquakes of 1928 July 18 and 1930 October 24. Bullen and I had to decide which of these should be regarded as a typical earthquake, and chose 1928 July 18, on the ground that the early S on 1930 October 24 might be due to focal depth. This, as it turned out, was wrong.

Beyond 83° SKS precedes S. It had given Turner much trouble. The practice at most stations was to read P and L first, L being the beginning of the surface waves. The interval between these gives a rough estimate of the distance, and hence of the time when S should arrive, within a minute or two. But beyond 83° an observer reading the first sharp movement about that time would inevitably read SKS for S. The result was that the I.S.S. contained hardly any readings of S at these distances, nearly all being SKS. A proper criterion for the separation was given by Lehmann and Plett. SKS, having been of P type in the core, is wholly of SV type ; but S at any distance includes SH. If a station has records of both horizontal components, they give the total horizontal displacement at each instant. Until S arrives this is in the same direction as for P, but the arrival of S is indicated by an abrupt change both in amount and in direction. Also the interval between P and SKS from 83° to 105° is nearly constant at about $10^m\,40^s$. It appeared, therefore, that if an observer finds an interval of about this amount between P and S he should suspect that the apparent S is really SKS, and should look for the true S after it. Since then the number of genuine S readings at large distances has greatly increased.

SKS resembled S in showing the Z phenomenon ; it and SKKS were reduced as far as possible to the same origin time as S.

For PKP the residuals showed a violently asymmetrical law of error. The normal law of error does not hold for any phase. The residuals usually show a concentration about a mode, this by itself suggesting a standard error of 2^s or 3^s ; but there is a long range on both sides within which the numbers of residuals fall off very slowly up to about 10^s. A mean and a mean square error found from

the residuals in the usual way would mean hardly anything. The explanation lies chiefly in a difficulty of identification. After a phase has passed, the ground does not return to rest, but continues to vibrate irregularly, usually with a diminishing amplitude, until the next phase arrives. If different observers agree on what to read they generally agree to 1ˢ or 2ˢ, but an observer reading his records without reference to the other stations is liable to take some part of the irregular movement as a genuine phase, and this is shown by a large residual. Even P does not usually start from rest, since there are perpetual vibrations of the ground, known as microseisms ; these lead to uncertainties about the position of the true beginning of P on the record and to large residuals of either sign. There is also occasional trouble from large clock errors. This was dealt with by detecting the offending stations and afterwards ignoring them. The result is that we have to deal with a modified law of error, with a sharp peak superposed on a nearly uniform background. I found that this can be dealt with without prohibitive trouble by the method of "uniform reduction", in which the residuals are grouped at 1ˢ intervals and counted ; then a constant is subtracted from the number in each interval, including the central ones, so as to leave the central group isolated, and a mean and standard error are found from the latter as reduced. Later work (1936) showed that this is a close approximation to the most accurate treatment, which supplies a weight to be attached to each observation as a function of the residual. This weight is approximately the probability that that particular observation is normal.

The usual methods, such as rejecting observations while treating those retained at full weight, are completely unsatisfactory. The decision about which observation to reject and which to retain is arbitrary, and may displace the mean by the full amount of its apparent standard error. If the study of the distribution of the residuals shows that the normal law of error is not followed, the correct procedure is to acknowledge the fact and devise a method of reduction suitable to the actual law. This is not prohibitively difficult in practice—certainly nothing like as difficult as making the observations (Jeffreys, 1938 b, 39 a).

For PKP there were a few early readings as for P, then a sharp rise to a maximum within a range of about 2ˢ, but then the readings dropped off very slowly over about 20ˢ. Bullen and I made a provisional solution, but without much confidence. The only other useful phase recorded with appreciable frequency in the I.S.S. was SKKS. The residuals for this were very scattered, but a solution had to be made because SKS and SKKS together make it possible to compute times of ScS without further assumption, and hence to infer times of other core waves to serve as standards of comparison.

As far as possible we relied on the reports in the I.S.S. for observational data. Numerous special studies of individual earthquakes had been made, but their authors usually appeared to disagree, and it was not possible to say how far the differences were due to some personal equation, whether some correlation between errors due to the method of reading might have led to an appearance of excessive accuracy, or whether there was some genuine difference between earthquakes. Using the I.S.S. we knew that all records for each earthquake were read by different observers, so that the personal equation could legitimately be treated as random error, and would be included in the estimate of uncertainty

derived from the scatter of the residuals. This method had the further advantage that the two most laborious parts of the work, the actual reading of the records and the preliminary calculation of an approximate epicentre and distances, had already been done for each earthquake by the stations and the Oxford staff. There were questions that could be settled only by special study, but this method indicated which these were.

This consideration is particularly important because a beginner in seismology, faced with an actual record, is often unable to distinguish at all between definite phases and irregular oscillation. This may lead to an attitude of complete scepticism. The reading of seismograms is an art that requires experience. Even experienced observers sometimes disagree ; but when a large number of them agree we can be sure that they are right. In near earthquakes Pg and Sg are frequently read in routine observations, and the results of special studies may be considered confirmed ; but P* and S* are seldom read, except for what appears to be S* in Japan, and to that extent it may be doubted whether they are genuine. Special studies serve a useful purpose in settling doubtful points, but the devising of adequate checks is more difficult than with routine observations.

The tables published by Jeffreys and Bullen (1935) were adopted in place of the Zöppritz tables for the I.S.S. reductions from 1930 January. The work since then has consisted mainly of tracing and eliminating systematic errors. The first of these was the effect of the ellipticity of the Earth. Gutenberg and Richter had suggested a method of allowing for it, but we were not satisfied that this was the best, and other outstanding errors seem likely to be larger, and their treatment could not be postponed until the calculation of the ellipticity effect was done in detail. The trouble was that the layers of equal velocity within the Earth would also be elliptical, and this would affect the times of transmission even between points at the same distance from the centre and from each other. Since Bullen's return to New Zealand at the end of 1933 he has calculated the correction for all the relevant phases—a most laborious piece of work—and allowance has been made for it. The final conclusion is simple (Bullen, 1936, 37 a, and later papers in same volume). If geocentric latitudes are used, the effect of ellipticity is allowed for within about $0^s{\cdot}2$ by adding a term proportional to the sum of the heights of the epicentre and the station above the sphere of equal volume ; the other factor is tabulated by distance for each phase. The form of the correction is more complicated if geographic latitudes are used. Bullen has dealt with this by providing a comprehensive table of triple entry for converting geographic to geocentric distances (1938). The difference can reach $0°{\cdot}4$, and in extreme cases an epicentre may be wrong by a quantity of this order if geographic distances are used in finding it. Incidentally this correction accounted for a number of cases where an earthquake in a medium latitude had given a set of very consistent negative residuals of -3^s or -4^s on the other side of the equator ; when geocentric latitudes were used these fell into line. The correction of the P table was simple because the earthquakes used fell into a few separate geographical groups : South-Eastern Europe and Western Asia ; North and Central America ; Japan ; South America ; and Pacific Ocean. For earthquakes of one group the ellipticity correction would be nearly the same, and its values for a mean epicentre could be applied at the end of the work. The corrected times for the groups

could then be combined, keeping the previous standard errors, since the correction was a known quantity. Thus the P table was reduced to a mean sphere (Jeffreys, 1937 c).

A further correction was needed for P at large distances (over 90°) and PKP, on account of the prevalence of late readings due to weakness. The trouble is that if they are small the observer usually waits till he is sure of them before he makes his reading, and the result is that the readings are systematically late. Up to about 90° there is little trouble because the law of error is nearly symmetrical. The difficult ranges were treated by using only the best stations with instruments for recording vertical movement—the latter because P at large distances and PKP rise steeply and are most prominent on the vertical. I found that this selection gave symmetrical distributions of residuals, with about the same standard error and background effect as for P at moderate distances, and the effect of skewness was satisfactorily eliminated. The greatest change from the Jeffreys and Bullen (J.B.) times due to correction of these two systematic errors was $-2^s{\cdot}7$.

For S and SKS the most satisfactory procedure was to start again, using geocentric distances throughout. The residuals were much more scattered in some earthquakes than in others, and in some cases seemed to fall into two or more distinct series separated by several seconds. A summary based on observations from two series would be representative of neither. At this stage, therefore, attention was confined to earthquakes with long series of S and SKS readings, consistent enough to indicate that nearly all stations over a long range of distance had read the same thing (Jeffreys, 1939 b, 39 c). To get a sufficient number of SKS observations at large distances it was necessary to use a number of earthquakes in the southern hemisphere (Jeffreys, 1938 a); it has been recorded up to 145°. When these data were combined it was found that the J.B. times of S beyond 80° and of SKS were about 12^s too long. This had already been inferred from some deep-focus earthquakes (Jeffreys, 1935 a), which will be discussed in a moment. On applying Pearson's χ^2 test to the mean residuals of these earthquakes over 5° ranges of distance it was found that they were as consistent as their apparent standard errors would suggest, so that the correctness of the revised tables is adequately checked.

§7. DEEP-FOCUS EARTHQUAKES

These were discovered by Turner in 1922, using the I.S.S. data. His method of determining a preliminary epicentre was to draw an arc about the point of a globe that represented each station with radius equal to the distance corresponding to the interval from P to S. These arcs usually intersected near one point and gave a rough epicentre. But he found that in some cases the arcs failed to meet by several degrees. The data could, however, be reconciled fairly well by supposing P and S to have come from a focus at a considerable depth. When this happened, if PKP was also observed, it was found to be systematically early. A criticism of Turner's method was that the discrepancies in time were of the same order of magnitude as the errors known to exist, though not then well determined, in the Zöppritz tables, and that the possibility that they arose from errors of the comparison tables could not be excluded.

K. Wadati (1928) studied the arrivals of P and S in Japan in local earthquakes, also relying principally on the interval between P and S. This provides an estimate of the distance from the focus. In a deep-focus earthquake it would be larger at the epicentre than in a shallow one, but would not increase so fast with distance from the epicentre. Wadati found substantial differences between earthquakes in this respect. Further, it happened several times that both he and Turner found about the same focal depth for an earthquake by their different methods. His tables were much more accurate than those of Zöppritz, but it appeared possible that some of the upper-layer waves might have been mistaken for P and S. Consequently both methods, though suggestive, were not quite convincing.

Decisive evidence was obtained by Stoneley (1931) and Scrase (1931). By a reciprocal theorem in the theory of small oscillations, a normal mode cannot be excited by an impulse at one of its nodes. Now a surface wave of given period is a normal mode, and effectively the whole region more than a wavelength deep is composed of nodes. Hence a deep-seated disturbance should give rise to no surface waves of lengths less than the focal depth and of appreciable amplitude. The theory was worked out in some detail by Lamb (1904) and Nakano. Now in the alleged deep-focus earthquakes the I.S.S. gave numerous readings in the column devoted to surface waves, which provided further reason for doubt. But Stoneley found that these readings were much earlier than in shallow earthquakes, the differences reaching 10 minutes. They agreed well with the times expected for various S movements (SS, SSS, etc.) reflected at the outer surface, which would exist for any depth of focus. When original records were inspected, the true surface waves were untraceable at the proper time for them. It appeared therefore that the deep foci did exist. Stoneley's work provided the first unambiguous evidence.

Scrase followed up an earlier suggestion of G. W. Walker. If the focus was in the outer surface, PP would be reflected midway between the focus and the observer. But if the focus is at some depth two phases of PP type are possible, one reflected about distance $\frac{1}{2}\Delta$, as in a shallow earthquake, but the other reflected close to the epicentre. The latter is actually the more familiar in the analogous case in optics, since it corresponds to the rays that form the image in a concave mirror. The phase reflected at half the distance is much less familiar in the optical case because the mirror seldom includes a sufficient amount of the sphere. Scrase denoted the near reflexion by pP, keeping PP for the distant one, and similarly he denoted the other possible near reflexions by sP, pS and sS. He found, on examining the Kew records of a number of deep shocks, that these reflexions were present and very clear. Many stations had read them for themselves and reported them as unidentified additional readings. The intervals between P and pP or sP, and between S and sS, should vary slowly with distance; this also was verified. The verification of these near reflexions provided another crucial test of the hypothesis of deep foci, and there was no further occasion to doubt their reality. Stoneley's test and Scrase's, it must be noticed, do not depend on having accurate times of transmission as a preliminary, and are independent of the possibility of errors of the types that cast doubt on the methods of Turner and Wadati.

When accurate times of transmission were available I found it possible to make Turner's method much more sensitive. An approximate calculation of the effects of focal depth made it possible to find focal depths from P alone. In the five earthquakes chosen (1935 a), if the focus was assumed shallow, the residuals of P varied systematically with distance by 14^s to 60^s, however the epicentre might be chosen, but with allowance for depth the standard deviation was reduced to about 2^s, which is as good an agreement as is usually found for a normal earthquake. Analogous results were found for S ; in fact S appears to be better observed in many deep earthquakes than in normal ones.

It is convenient to record focal depth below the top of the lower layer as a fraction of R, the radius of that layer. The first work that indicated to me (1935 a) that there was something seriously wrong with the J.B. times of S at large distances and SKS was on the Afghanistan deep-focus earthquake of 1929 February 1. A long series of European stations had read S, and many North American ones SKS. The former gave small residuals, the latter very consistent ones of about -12^s. The only possible explanation seemed to be that the J.B. times of SKS, and, therefore, of S at large distances (with which it had been accurately compared in the process of identification), were about 12^s too long. This was confirmed by three other well-observed shocks, the deep ones of 1928 March 29 and 1931 February 20, studied by Stechschulte and Scrase, and the Marianne Islands one of 1930 October 24, studied by Lehmann and Plett. We had actually used Lehmann and Plett's readings in the latter, but they all referred to great distances, and the comparison with intermediate distances became possible only when the data had appeared in the I.S.S. Scrase had actually remarked that $dt/d\Delta$ for S beyond about 70° in his study (1933) was less than my 1932 times gave, but the J.B. values are larger than my 1932 ones. I was not sure then that the difference was not due to some error in estimating the allowance for focal depth, but later work has fully confirmed it.

Deep-focus earthquakes have also been useful in solving the problem of the times of S up to 20° to 25°. S can be observed at these distances in normal earthquakes if attention is paid to the change of direction of the movement, but it is followed by several large movements, one or other of which is read by most stations. No satisfactory explanation has been given of these movements. It is, however, clear that they are not S. In deep-focus earthquakes, however, the S residuals at these distances show a pronounced concentration, with only the usual amount of background effect. A study of ten such earthquakes in and near Japan, combined with the velocity at short distances found from European near earthquakes, has led to what appears to be a highly satisfactory determination of the times of S up to 20°.

Another question settled by deep-focus earthquakes is whether the interfaces postulated in near earthquake study are true discontinuities or continuous transitions. Near earthquakes by themselves can only establish the existence of approximately uniform layers, and say little about the transitions. But if there are discontinuities, the usual pP, reflected at the outside, should be preceded by a smaller reflexion from each interface. Direct examination of seismograms of deep shocks showed no such reflexions, and, therefore, the transitions must be continuous (Jeffreys, 1935 a). It is, however, still necessary to use sharp transi-

tions in any model used for calculation, because we have still no means of estimating how thick the zones of transition are, and an over-estimate of their thicknesses would be as bad as treating them as zero, and more difficult to correct.

If we adopt a standard structure of the upper layers we can find the focal depth from the observations of P and S at different distances. But given this depth we can also calculate the times of pP and sS. These are more sensitive to the thickness of the upper layers than P and S are, because they have passed through them three times instead of once. Consequently the observations of pP and sS provide an additional equation of condition for the thicknesses of the upper layers. By combining this with the surface-wave data I get 15 ± 3 km. for the upper layer, 18 ± 4 km. for the intermediate one. With these values it is possible to find times corresponding to a surface focus, which is the most convenient standard of comparison.

The core reflexions PcP and ScS are much easier to read in deep shocks than in normal ones, chiefly on account of the smallness of the surface waves. The readings of PcP are still rather scattered, presumably because a P wave striking the core is largely transmitted. ScS is quite satisfactory, and leads to the best determination of the radius of the core.

§8. THE 20° DISCONTINUITY

Byerly's work on the Montana earthquake and my earlier work showed that $d^2t/d\Delta^2$ for P is very large about distance 20°. The same had already been noticed by F. Neumann, and emerged again in the earlier stages of my work with Bullen. At short distances $dt/d\Delta$ is about $14^s{\cdot}3/1°$, decreases gradually to $12^s{\cdot}3/1°$ between 18° and 19°, and then drops to $10^s{\cdot}4/1°$ between 20° and 21°. At larger distances it falls off steadily, reaching $4^s{\cdot}4/1°$ beyond 100°. These facts indicate that between the depths reached by the rays that emerge at 19° and 20° there is a great increase of the velocity of P. There might be a new discontinuity, which was supported by some observations made by Lehmann (1934). She found what appeared to be a second P in records between 20° and 24°, just as Pg in near earthquakes continues to be readable in near earthquakes at distances where it is preceded by P. Her readings, however, are a few seconds too late to lie on a smooth continuation of the P curve up to 19°. The large amplitudes of P about 20° also suggest a strong curvature of the time-curve and, therefore, a rapid continuous increase of velocity with depth above and below the discontinuity, if there is one. The question was, therefore, not whether there are two regions, within each of which the velocity varies nearly uniformly with depth, and separated by a sharp discontinuity, but whether the increase is or is not sufficiently rapid to make the distance reached by a ray less for some rays than for some others that have begun by descending less steeply. If it is not, the time-curve is a smooth curve with a strong curvature about 20°; if it is, the curve has a loop terminating at each end in a cusp, and for distances within this range there will be three P's. The oscillations that follow P everywhere make it impossible to identify these with confidence. Consequently the question could not be decided by means of normal earthquakes alone. It was tested by means of the Japanese deep shocks (Jeffreys, 1939 b). Separate calculations of

P times were made for different focal depths, one assuming everywhere finite curvature of the time curves, the other assuming a sharp discontinuity. They differed by amounts up to $1^s{\cdot}4$ for the depths of these earthquakes. If the difference between the times of the first P on the two hypotheses is x, and the true time differs from those on the hypothesis of discontinuity by ax, a could be estimated from the data, and was found to be $+0{\cdot}17 \pm 0{\cdot}21$. This amounted to a disproof of the first hypothesis, which would correspond to $a=1$. Nevertheless the second is not altogether satisfactory because it does not explain the large amplitudes. Finally, a continuous distribution of velocity was adopted, the variation over a limited range being rapid enough to give a triplication of P at the corresponding distances. In the extreme case of a true discontinuity, the receding branch would reduce to the pulse reflected at it ; this was sought but not found. In deep shocks the first arrivals of P and S at some distances corresponded to parts of rays that did not give first arrivals in normal shocks, and consequently part of the range where P and S are triplicated could be reconstructed from the data for the deep shocks. Some arbitrariness remained, but at any rate the solution reconciled the whole of the data.

§9. THE INNER CORE

On the simplest theory, corresponding to geometrical optics, there would be no P or PKP between 105° and 142°. Actually a feeble extension of P from 105° exists, and has been traced as far as 150°. This is due to diffraction around the core. It is quite different in appearance from P proper, having no sharp beginning. From PKP at 142°, again, an extension is traceable back to about 110°. This was originally interpreted as also due to diffraction, in this case near a caustic. Lehmann (1936) and Gutenberg and Richter (1938, 39), however, rejected this explanation, because it did not account for the short periods and large amplitudes to be found at these distances. I found (1939e), by applying Airy's classical theory of diffraction near a caustic, that periods of the order of 1^s should be traceable back to about 139°, and those of 10^s to about 128°. Actually periods of 1^s, associated with large amplitudes, are observed between 123° and 125°. The diffraction theory was, therefore, impossible, and it was necessary to construct one of refraction. But to get a branch of such length requires very exceptional conditions. A reflected branch would always be associated with an earlier refracted one. The deep earthquake of 1932 January 9 gave a fine series of observations. It was found that the time-curve is nearly straight, and there was considerable difficulty in framing a hypothesis that would give the observed length. A sharp increase of velocity with depth, well within the core, would give the longest reflected branch with any ordinary hypothesis, but calculation showed that it could not be long enough unless, in addition, the velocity decreased with increase of depth for some distance outside the discontinuity. Rays grazing what we must call the inner core could be reflected ; those striking it sufficiently steeply to penetrate would give a refracted branch giving slightly smaller times of travel, for equal distances, than the reflected wave. On this hypothesis the data could be fitted (Jeffreys, 1939f).

§10. THE VELOCITY DISTRIBUTION

If the times of a pulse satisfying the principle of stationary time are known over a range of distance starting from 0°, the distribution of velocity can be found by a series of straightforward numerical integrations. Theoretically this remains true even in cases of strong refraction leading to a triplication of the pulse over some range of distance, but in practice the upper branches are unobservable, and some hypothesis is needed to reconstruct them. For a shadow zone the pulse does not exist at all over a range of distance, and the method fails. It also fails, in theory, if the times are not known over a range near the origin, but in that case they can usually be inferred without difficulty. It can be shown that if the velocity has a continuous derivative with regard to the depth for small depths, the times of arrival have the form $\alpha\Delta-\beta\Delta^3+\ldots$, and if we can find distances where the time is known, and this formula agrees with them, the formula can be used to infer times for all shorter distances.

If a pulse undergoes reflexion or discontinuous refraction at any stage we can make direct use of the principle of stationary time. To take a simple example, that of PS, let the point of reflexion be at distance Δ_1 as seen from the centre of the Earth and the time of P at it be T_1. Then the angular distance travelled as S is $\Delta-\Delta_1$, and S travels this distance in time T_2. The total time is then T_1+T_2, and for it to be stationary for small variations of Δ_1, for fixed Δ, we must have

$$\frac{\partial T_1}{\partial \Delta_1}=\frac{\partial T_2}{\partial(\Delta-\Delta_1)}=\frac{\partial(T_1+T_2)}{\partial \Delta}.$$

To construct a table for PS, therefore, all we have to do is to inspect the tables of P and S to find distances where the values of $\partial T/\partial\Delta$ are equal ; and then, by adding the distances and times, we get a distance for PS and the time corresponding to it. Repeating the process for different values of $dt/d\Delta$ we can construct the complete table for PS. In practice an error in Δ_1, for given Δ, makes only a second-order error in the sum T_1+T_2, and great accuracy in matching the gradients is not necessary ; tabular entries at 1° intervals are quite enough. Calculated times to 0ˢ·1 are necessary to prevent accumulation of rounding-off errors ; the actual standard errors are mostly of 0ˢ·3 to 0ˢ·5 for the well-observed phases.

Conversely, if we had the times of P and PS, but not of S, we could infer those of S by matching gradients and subtracting distances and times.

This process can be used to infer times within the core. If we have times of ScS and SKS with the same $dt/d\Delta$, the time spent by SKS within the outer shell is that of ScS, and by subtraction we get a corresponding time and distance for the wave in the core alone. Similarly we can infer times of K from those of PcP and PKP ; the range of Δ in the core is not the same in the two cases, the first case giving the smaller distances and the second the larger, with a slight overlap. The distances for K do not come down to zero, but some of them are small enough for a cubic formula to be fitted to the times, and thus the times can be calculated for short distances between points on the surface of the core. The removal of the effects of an outer layer in this way may be called "stripping". It is applied in succession to remove the effects of the upper and intermediate layers, then the

outer shell, and finally the outer part of the core, leaving the inner core, which, as far as we know, contains no further discontinuity.

Typical velocities in km./sec. found in this way (1939e, 39f) are as follows:

		P	S
Upper layer		5·58	3·36
Intermediate layer		6·50	3·74
Lower layer :	top ($r=R$)	7·747	4·35
	$r=0{\cdot}94R$	8·97	4·96
	$r=0{\cdot}93R$	9·50	5·23
	$r=0{\cdot}56R$	13·64	7·31
Core :	$r=0{\cdot}548R$	8·10	
	$r=0{\cdot}230R$	10·37	
Inner core :	$r=0{\cdot}20R$	11·16	
Centre :	$r=0{\cdot}00R$	11·31	

It is convenient to use R, the radius of the base of the intermediate layer, as the standard of length. For deep-focus earthquakes the depth below the base of the intermediate layer is given as a fraction of R.

In Gutenberg's original method the depth of the core was found as the greatest depth reached by P before the shadow begins. This was not very satisfactory because the shadow is not sharp, and the determination requires the slope of an empirical curve at its very end, always a matter of considerable uncertainty, but at that time no other method was possible. We can now determine the velocities of P and S, and extrapolating them slightly we can compute times of PcP and ScS for trial depths of the core; then comparison with the observed times gives an estimate of the depth of the core. The radius of the core is found to be $(0{\cdot}5480 \pm 0{\cdot}0004)R = 3473 \pm 2{\cdot}5$ km. With this datum the times of PcP and ScS are computed, and are used in determination of times in the core.

On any admissible hypothesis, a reflected pulse from the outside of the inner core follows too quickly after the refracted one to be read separately, and the determination of the radius is less definite.

The final stage is the computation of the time of various compound phases, especially those that have undergone one or more reflexions at the outer surface or the core boundary. This has been done (Jeffreys, 1939f; Bullen, 1939b) not only for a surface focus but for depths at intervals of $0{\cdot}01R$ to $0{\cdot}12R$. The greatest focal depth known so far is about $0{\cdot}093R$ (originally reported as $0{\cdot}11R$, but this turned out to be too great when more observations became available (Jeffreys, 1941)). A large number of compound phases are readable, but in general with less accuracy than the main phases P, S, PcP, ScS, PKP, SKS, and the deep-focus phases pP, sS.

In a long series of papers over about the same period as mine and Bullen's, Gutenberg and Richter have made independent determinations of the times of the principal phases and the velocity distributions. On the whole their results agree well with ours, and we were often enabled to settle doubtful points quickly by consultation with them.

§11. IDENTIFICATION OF MATERIALS AND DENSITY DISTRIBUTION

Much experimental information is available about the mechanical properties of rocks at pressures equivalent to those existing at depths up to 60 km. This

range covers the depths relevant to near earthquakes. The most satisfactory method of comparison seems to be to use the equation

$$\alpha^2 - \tfrac{4}{3}\beta^2 = \frac{\lambda + \tfrac{4}{3}\mu}{\rho} = \frac{k}{\rho}.$$

The quantity on the left is determinate from the velocities alone, and for any material tested the right side is the ratio of the bulk-modulus to the density. The bulk-modulus is the easiest elastic constant to measure at high pressure, and in particular has a negligible time-effect. It was originally expected from geological considerations that the upper layer would be granite and the lower "basalt", the latter for this purpose including the crystalline forms dolerite and gabbro. Suess's terms "sal" (altered to "sial" by Wegener) and "sima" were originally intended to allow somewhat greater latitude of interpretation, but in practice have come to be used in the restricted senses. The following table compares measurements by L. H. Adams, E. D. Williamson and R. E. Gibson at the Geophysical Laboratory, Washington, with the seismological data.

	$\alpha^2 - \frac{4}{3}\beta^2$ (km./sec.)^2		k/ρ (km./sec.)^2	ρ (at 2×10^9 dynes/cm^2)
Upper	16·1	Granite	18·3–19·5	2·61
Intermediate	23·6	Diorite	22·8–24·3	2·74
Lower	34·8	Gabbro	27·8–28·2	3·06
		Tachylyte	24·2	2·85
		Dunite	35·2–38·3	3·29

The order of increasing k/ρ in the crystalline rocks tested is that of increasing density and that of decreasing silicon and alkali content. Tachylyte is the glassy form of basalt. No rock tested gives a higher value of k/ρ than dunite, which is nearly pure olivine $(Mg, Fe)_2SiO_4$. This value agrees very well with the seismological value for the lower layer. The seismological value for the upper layer is rather low for a typical granite; perhaps the most likely interpretation is that the upper layer contains more free silica than the average granite, or it might contain some of the glassy form, obsidian. No layer agrees reasonably well with gabbro, though the intermediate layer would agree with either tachylyte or diorite. According to modern petrological theories, igneous rocks are considerably altered in composition by partial crystallization in rising from their original levels to accessible places. Then we should have to regard the typical surface granite and basalt as having come from rather low down in the upper and intermediate layers respectively, possibly from the transition zones. This is reasonable also because in each layer the temperature is nearest the melting point at the bottom. On account of the close correlation between k/ρ and ρ, if we are concerned with properties that involve only the density, the latter is well determined for each layer, since all ordinary rocks that give the same k/ρ have nearly the same density.

Two other data relevant to the distribution of density are the mean density and the moment of inertia. The latter is very accurately known for the following reason. If C is the moment of inertia about the polar axis, A that about an axis in the equator, the rate of the precession of the equinoxes contains a factor $(C-A)/C$, the other factors being very accurately known. If E is the Earth's mass and a its equatorial radius, the distribution of gravity with latitude and three measurable perturbations of the moon depend on $(C-A)/Ma^2$. The ratio of

these numbers determines C/Ma^2, which is $0{\cdot}3345 \pm 0{\cdot}0012$. For a homogeneous sphere this ratio would be 0·4, so that we can infer that the Earth has a considerable central condensation. Various distributions of density have been chosen to fit these (or equivalent) data. A more natural one was due to Wiechert in 1897, who guessed that the Earth might consist of a rocky shell and a metallic core, and supposed both uniform in density. Such a constitution involves three unknowns, the two densities ρ_0, ρ_1, and the radius of the core, $a\alpha$, one of which can therefore be chosen arbitrarily. He took the plausible value 3·20 for ρ_0, and got $\rho_1 = 8{\cdot}22$, $\alpha = 0{\cdot}78$ (with a slightly different mean density from that now accepted). But seismology shows no discontinuity anywhere near the boundary of Wiechert's core. If on the other hand we take the seismological value $\alpha = 0{\cdot}545$, we get $\rho_0 = 4{\cdot}22$, $\rho_1 = 12{\cdot}32$. The explanation of the discrepancy between ρ_0 and the density of dunite is simple ; each layer of the Earth is compressed by the weight of the material above it, and, therefore, even a layer uniform in composition would have a density increasing with depth. A rough comparison showed that if the material of the shell had density 3·2 gm./cm^3 at atmospheric pressure, the actual mean density of the shell would be in the neighbourhood of 4·2. Similarly if we regard the core as liquid, the velocity of longitudinal waves in it gives k/ρ directly ; assuming a mean density of 12 gm./cm^3 at the actual pressure, and then supposing this pressure removed, I found (1926) that the density would sink to about 8 gm./cm^3, about the density of iron. Consequently Wiechert's principle was valid, and the rough agreement between his densities and those of dunite and iron persisted ; his radius for the core was wrong because he had not allowed for compression. He wrote, of course, before P and S had been identified, much less PKP, so that he had no geophysical data for estimating the compression.

The hypothesis that the core is liquid is strongly suggested by the failure to identify S waves in it. This does not quite settle the matter, because they might exist but be strongly damped. There is, however, much confirmatory evidence. Kelvin showed long ago that the tidal yielding of the earth indicated a rigidity comparable with that of steel. But the rigidity of dunite at atmospheric pressure is about two-thirds of that of steel, and it follows from the increases of the density and the velocity of S with depth that at the base of the shell the rigidity is about three times that of steel. A calculation (Jeffreys, 1926) showed that if the core was taken as liquid we could get a good agreement with the tidal yielding ; if it was taken as having a rigidity comparable with its bulk-modulus the tidal yielding would be far too small. The core therefore certainly behaves as liquid under stresses with periods of the order of 12 hours. Even this does not quite close the matter, because a solid might have such imperfection of elasticity that it would behave as liquid to tidal forces but as solid to stresses with periods of a few seconds. This not very hopeful suggestion, however, is hard to reconcile with the size of SKS. Many cases of refraction at interfaces between solids have been worked out, and one peculiarity is common to all so long as the velocities of P and S are in about the same ratio ; an incident P gives a very small transmitted S, and conversely. But SKS has undergone two such transformations and yet is comparable in size with S. It seems, therefore, that any escape from the conclusion that the core is liquid would have to be wildly artificial.

We have now much more detailed information about the distribution of density from the work of Bullen (1936, 37 b, 40 a–41). This started as a step in the work on the ellipticity correction in seismology. The ellipticities of internal surfaces of equal density were required, and a preliminary stage was the determination of the density-distribution itself. This requires the solution of the differential equations

$$\frac{dp}{d\rho} = \frac{k}{\rho} = \alpha^2 - \tfrac{4}{3}\beta^2 ;$$

$$\frac{dp}{dr} = -g(r)\rho,$$

where r is the distance from the centre, p the pressure, and $g(r)$ the local intensity of gravity, itself depending on the values of ρ for smaller values of r. These can be solved numerically, the seismological values of α and β being used for each r. After allowing for the outer layers, Bullen proceeded inwards to the core boundary ; but he found that any density just within the core that would lead to the correct mass would give too small a moment of inertia. This approach, in fact, had assumed only one adjustable constant, and another was still needed. But there was a possibility of another discontinuity of density, namely at the 20° discontinuity. He took this to be a sharp change at a depth of 350 km., and found that with a jump of density as well as of the earthquake wave-velocities he could get an agreement. I found a little later that if the change is treated as a true discontinuity its depth is about 480 km. ; with this modification the jump of density is from 3·68 to 4·20 gm./cm³. If the pressure was removed from the lower material its density would be about 3·8 and the velocity of a P wave in it about 8·4 km./sec.

After inspection of the data for some likely materials I found (1937 b) that the only one with about the right properties was MgO. A most interesting alternative suggestion was made by J. D. Bernal. The olivine crystal is a rhombic lattice of oxygen atoms with the silicon and metallic atoms in a somewhat asymmetrical pattern. At high pressures there is a possibility of a complete rearrangement, suggested by the corresponding germanium compound Mg_2GeO_4, which has been made by Goldschmidt, and is isomorphous with Mg_2SiO_4. Ge has larger atoms than Si, and if there is any question of a rearrangement under high pressure the Ge atoms would obstruct the further compression of the oxygen lattice at smaller compressions than Si atoms would. Now a denser form of Mg_2GeO_4 exists at ordinary pressures, its form being cubic. It is therefore plausible that at high pressures olivine may change over from a rhombic to a cubic form with a sudden increase of density, and that this is the explanation of the 20° discontinuity.

The pressure at a depth of 480 km.* has not yet been imitated in the laboratory (Bridgman has attained that at a depth of about 150 km.), but a striking confirmation is found by comparison with the Moon (Jeffreys, 1937 b). If the material between the 20° discontinuity and the core is a new material it occupies about 80 per cent of the volume, and we should expect it to compose also most of the Moon. But the Moon's mean density is only $3{\cdot}334 \pm 0{\cdot}003$ gm./cm³, prac-

* 16×10^{10} dynes/cm²

tically that of ordinary olivine. If it had the composition of the Earth's shell, on the hypothesis of a new material, the density would be about 3·8 gm./cm.^{3} Consequently this hypothesis meets with formidable difficulties if we are to suppose any sort of analogy between the Earth and the Moon. On the other hand the pressure at a depth of 480 km. in the Earth is not reached at all in the Moon and, if the 20° discontinuity is a pressure change, the corresponding increase of density will not be present in the Moon, and the mean density of the Moon is explained without any further hypothesis. This check strongly confirms Bernal's suggestion that the material below the 20° discontinuity is cubic olivine. Further, if the Moon is uniform in composition (apart from a slight correction for the outer layers) we can calculate what the density of the main material would be at zero pressure ; it would be 3·29 gm./cm.^{3}, in perfect agreement with the standard specimen of dunite used by Adams and Gibson.

The material is presumably not pure, so that there would be no ground for surprise if the discontinuity is smudged out to some extent ; the evidence that it is a rapid transition rather than a sharp one does not affect the main argument. Bullen has taken the matter considerably further in later papers (1942a, 42b).

The inner core remains a problem. Its volume is so small that any reasonable density for it would hardly affect the rest of the results.

From analogy with meteorites, Goldschmidt had inferred that a layer of metallic sulphides should surround the core. This would be liquid and would probably be immiscible with liquid metal. Consequently there should be independent reflexions at its top and bottom and ScS should be double. This is not so. I also found serious difficulties in fitting the times of K at short distances on this hypothesis (1936). It appears that if a sulphide layer exists it cannot be much more than a tarnish.

§ 12. OUTSTANDING PROBLEMS

It became clear at an early stage of the revision of the tables that the chief need if any substantial progress was to be made was for better methods of combining observations than had hitherto been used. Consequently the seismological work has been carried on simultaneously with investigations into the principles of scientific method and the development of probability theory (Jeffreys, 1939). Considerable advances had been made within the present century, especially by Karl Pearson and R. A. Fisher, mostly for biological purposes, but seismology was found to present a number of problems that had not been treated (not counting the treatment of serious departure from the normal law of error, the principles of which had been given by Fisher). The most frequent of these was perhaps that of smoothing an empirical table when the true values were not expected to agree with any pre-assigned mathematical function ; the observed values would necessarily show some irregularity on account of observational error, and the question was how far this could be smoothed out without at the same time smoothing out some genuine irregularity. The solution of the problem should have much more general application. The principles of significance tests have had to be considerably extended because, whereas biologists have usually been able to design their experiments so as to

make a comparatively simple analysis of the results adequate, such design is impossible in seismology, and approximations have had to be developed. These may be expected to be useful in such matters as the revision of the fundamental constants of physics. Much more, however, remains to be done.

It was found that many of the differences that had been found previously disappeared on statistical examination, being found to be reasonably attributable to known random errors. Some remained; some of these have been explained, but not all. The Z phenomenon, an apparent variation in the times of S and SKS relative to P when different earthquakes were compared, gave some trouble at the start, but most of it is now explained. The correction for ellipticity accounted for several instances of it, and multiplicity for many others. Turner had found a large number of earthquakes in which the S-P interval for given distance was larger than in the normal earthquake, and attributed these to foci higher than normal. This led to the conclusion that the normal earthquake has a focal depth of at least 100 km.; but this was contradicted by the Montana earthquake, which certainly had an upper-layer focus and gave transmission times agreeing with Turner's average. The phenomenon was ultimately explained by Stoneley (1939). Some earthquakes are double or triple, the original one being followed by one or more larger ones after a few seconds to several minutes. An observer will naturally read the P of the first shock, since this is the first displacement on his record. But the most conspicuous S will be one of the later ones, so that the interval between the shocks is included in the observed interval between P and S. In practice many observers read two or more P's and two or more S's, and on comparison it is found that the readings of P and S can each be analysed into two series, separated by intervals independent of distance. There are, however, more readings of the first P and the second S, and if attention is confined to these, the phenomenon found by Turner will arise. Tillotson, by study of the additional readings reported in the I.S.S., showed that they could be explained in accordance with Stoneley's suggestion. This accounted for the other cases of positive Z that were known. A few cases of negative Z were probably due to focal depth, the near stations having been too few to enable it to be detected from P alone.

Even after the bulk of the Z variation had been explained in these ways, it was found that a little survived in deep-focus earthquakes. By a suitable choice of epicentre and focal depth, it was possible to fit the times of P at all distances, pP and sS, S at short distances and SKS within their apparent uncertainties; but the mean residual of S from 20° to 90° fluctuated from earthquake to earthquake as if it had a standard error of about $1^{s}\cdot5$ superposed on the random error. No explanation has been suggested of how such a variation, affecting only part of the range of one phase, could arise. There is a similar fluctuation in the difference of the constant terms in the times of the near earthquake phases (Jeffreys, 1937*d*); for this several explanations have been suggested, but none appears satisfactory. There are small but apparently significant differences in the times of P for earthquakes in different regions, the largest being in North America where they are about 3^{s} longer about 20°, in comparison with short distances, than in Europe and Japan (Jeffreys, 1940). It would hardly have been expected that the Earth's materials would have exactly the same properties everywhere; it is rather remarkable that they are as uniform as they are.

The greatest outstanding difficulty is one that possibly escapes attention because it is so familiar ; it is nothing less than the general appearance of the seismogram itself. According to Lamb's theory (1904) for a uniform crust, P, S and the Rayleigh waves from a sudden shock would all consist of single swings. The actual movement is a continual irregular oscillation. We know why the Rayleigh waves and surface waves of Love's type show oscillation ; it is due to dispersion. But no cause of dispersion has been suggested that is capable of converting a sudden P or S into a long train (Jeffreys, 1931 a, 31 c). The most hopeful at one time appeared to be internal reflexion in the upper layers, but this failed when reflexions at their interfaces were sought, but not found, in deep-focus earthquakes. If the original disturbance at the focus was oscillatory it would explain oscillations in P and S ; but the duration of the oscillation would be the same at all distances, whereas it increases with distance. This applies, in particular, to the suggestion, sometimes made, that the oscillation is due to non-linearity in the stress-strain relations near the focus.

A possibly related fact is that oscillation of the ground in a near earthquake continues long after the time that the slowest surface wave should take to arrive.

There is evidence that all seismic waves show damping as they advance. This is required in particular to explain why the upper and intermediate-layer phases become too small to be read at distances of 8° at most. The most likely explanation is scattering (Jeffreys, 1937 d, p. 220). If a layer is not quite uniform the rays will be bent irregularly and a sharp wave-front will be blunted. It is found that the time of most rapid increase of displacement will still be proportional to the distance, and its velocity will represent the average velocity in the material. There will be no sharp beginning, but a small diffuse movement will precede the main shock ; and it will be the main shock that obeys the fundamental rules of transmission. This seems to explain a curious anomaly found in near-earthquake studies. When a cubic formula is fitted to the times of P up to 20°, it gives $dt/d\Delta$ at short distances corresponding to a velocity of about 7·8 km./sec., which agrees with the value derived from most near-earthquake observations up to about 1930. Since then many studies have given velocities of 8·0 to 8·2 km./sec., which would not agree with the times up to 20°. Conrad, in his study of the Schwadorf earthquake, read two separate P's at the beginning of the records of some stations, the faster giving $8{\cdot}10 \pm 0{\cdot}09$ km./sec., but it was feeble, and not read at the majority of the stations ; while the later and larger one corresponded to a velocity of $7{\cdot}85 \pm 0{\cdot}04$ km./sec. The natural interpretation is that the second is the main pulse and the earlier a precursor due to scattering, and recorded only on instruments with very high magnifications. If so these are not an unmixed blessing unless their user is aware of the danger.

Another peculiarity is in the amplitudes at short distances. These decrease with distance, very roughly like $1/\Delta$, but definitely not so fast as $1/\Delta^3$. For smooth interfaces, theory indicates that the amplitude of Pg should decrease like $1/\Delta^2$ and P should *increase* slowly. There seems to be some possibility that theory and observation may be reconciled here if the interfaces are irregular, giving a diffusive effect like that of ground glass.

The travel time of any compound phase is stationary for small variations of the path, but not necessarily a minimum for all. It is easy to see that the time of

PP in a homogeneous sphere would be a maximum for variations of the point of reflexion in the plane of the path. Attempts were made to derive empirical times for all predicted phases, but it was found that there was a qualitative difference between those whose times were true minima and those whose times were merely stationary. For the former, namely P, PKP, S, SKS, PcP, ScS, pP and sS, the distributions of residuals showed pronounced humps corresponding to a standard error of 1ˢ to 3ˢ, with the usual background effect. The latter, typified by PP, PS, SS, SKKS and SKSP, give distributions scattered over 20ˢ or more with no pronounced concentration, and were finally abandoned ; their published times are calculated, the observed times being scattered on both sides of the calculated ones. Several attempts were made to construct a theory of a pulse of this class, but the mathematical difficulties were very great, and finally I adopted a suggestion of Mr. F. G. Friedlander, who pointed out that in the expansion of a cylindrical pulse the disturbance from the rear of the cylinder has the same property. An analytical solution of this problem was known, and there was no great difficulty in approximating to it. The result was that the disturbance should be very large, but not sharp ; it should begin a little before the theoretical time and subside gradually after it (Jeffreys, 1942a). Consequently we have at least a clue to, though perhaps not a complete explanation of, the unsatisfactory character of these phases.

The large scatter of the readings of S in normal earthquakes up to 20° is still unexplained ; it is clear that the large movements read as S by most observers are not S, but we still do not know what they are.

§ 13. LOCAL EFFECTS OF EARTHQUAKES

The instrumental recording of earthquake waves helps appreciably towards understanding what goes on near the epicentre of a large earthquake. The original movement is a fracture, in many cases a slip on a known fault. The records show definitely that the motion produced is a vibration, beginning with Pg and followed by Sg, the interval at short distances being a few seconds. Sg in general is larger than Pg and, since the SH component is usually the larger, may be nearly at right angles to it. It follows that buildings that withstand Pg may be wrecked by Sg, and if so they will fall in a completely different direction. In some earthquakes movement from different directions has been detected from the fall of buildings, and it has been inferred that there were really two earthquakes in different places ; actually the phenomenon is just what we should expect a single earthquake to produce.

At greater distances the change in the direction of the movement can be felt ; I noticed it at Cambridge in the North Sea earthquake of 1931 June 7.

Observers have reported that they saw lights during an earthquake, and there is a frequently quoted report that a man saw the waves travelling across country "faster than he could walk, but not so fast as he could run" (cf. Heck, 1936; Davison, 1936.) The velocity of Sg is about 3·3 km./sec., and the train of waves following it dies down so that the tail travels roughly at 1·5 km./sec. This observer should clearly have entered for the Olympic Games. But what we do know is that an observer in such conditions is undergoing horizontal and vertical accelerations applied to his feet, and must be making muscular efforts to keep his

balance. It would be a problem for an experimental psychologist to say whether in these conditions he would be likely to imagine that he sees a wave-like motion of the ground or to "see stars". A characteristic "earthquake noise" is also reported, somewhat resembling thunder, and the nature of this also needs investigation.

Pillars have often been found to have rotated during an earthquake. This has been attributed to rotation of the ground, but the prevailing wave-lengths are of the order of some kilometres. It is more likely that the rotation is due to the point of support, when detachment is imminent, not being vertically below the centre of mass.

It appears that an acceleration of $0{\cdot}1g$ is enough to produce almost complete destruction of buildings. Several non-quantitative scales of intensity have been produced, and are useful in the approximate location of epicentres; they have been calibrated quantitatively by Sieberg (1932), and a scale for estimating the total energy of an earthquake from instrumental data is given by Richter (1935).

The energy liberated by a major earthquake (one recorded at distances of 80° or more) is of the order of 10^{21} ergs. The small Hereford earthquake of 1926 liberated 5×10^{16} ergs and the Jersey earthquake of the same year 10^{19} ergs. The total displacement of the ground during a large earthquake may reach several metres; this may really be accumulated during several fractures at intervals over a few minutes.

A large earthquake is often preceded by foreshocks, and is nearly always followed by a series of aftershocks, diminishing in frequency and intensity. The former are probably slips on a small part of the fault-face where adhesion is relatively weak, the main shock coming when the stresses have grown sufficiently to produce general fracture. The aftershocks are readjustments following the redistribution of the general stress made by the main shock. There is good reason to suppose that the huge displacements shown in major faults (which may reach thousands of feet) are the result of many successive earthquakes at considerable intervals of time (Jeffreys, 1942b).

The disturbance felt in deep-focus earthquakes is more widely distributed than in shallow ones and less intense at the epicentre, for the same total energy, and this property often provides the first indication of focal depth.

The existence of deep foci is a serious difficulty for any of the hypotheses that require continuous flow at great depths to play a constructive part in geological phenomena. The tendency of continuous flow is always to reduce the local stress-differences; it therefore makes it more difficult to accumulate the large stresses shown to exist by deep earthquakes. It is possible that continuous flow does occur; but if it does, any discontinuous changes are in spite of it and not because of it. I have always preferred to work on the hypothesis that the materials behave as perfectly elastic until the stress-difference reaches a definite non-zero value, and then break suddenly. If this hypothesis cannot explain the facts, then, *a fortiori*, no hypothesis involving continuous flow can do so.

§14. SEISMIC SURVEY

The methods of the study of near earthquakes have been extensively applied to the investigation of the sedimentary layers. The stimulus probably came from the Oppau explosion of 1921, which was well recorded by seismographs to a distance of about 350 km. The waves are generated by firing a charge of high explosive, which must be buried to prevent nearly all the energy from going into an air-wave. If this is done there is hardly any change in the form of the surface afterwards. The position of the focus is known, and the time of the explosion can be recorded directly on the seismograms, so that the time of transit of each pulse is measured directly ; thus one of the chief difficulties of the study of natural earthquakes, the need to determine the epicentre and time of origin from the times of arrival themselves, does not arise. Further, the only way of making sure of getting a record of the natural earthquake is to keep the seismograph running all the time ; to do this, and keep the apparatus manageable in size, limits the speed of the paper to about 1 mm./sec., and usually much less. But in recording an artificial earthquake we are free to start the drum whenever we like, and consequently can run it much faster. It is therefore as easy to read the records to $0^{s}{\cdot}001$ as to 1^{s} in a natural earthquake, and it becomes possible to determine not only the mean depths of layers but variations about the mean. On the other hand the energy available is small compared with that in a natural earthquake. It would take several tons of explosive to match even the small Hereford earthquake, and the larger the charge used the more deeply it must be buried to prevent it from blowing out and wasting most of its energy in the air. It is probable that in some large quarry blasts Pg has been recorded, but the waves in the lower layer have never been.* In practice, therefore, the method is usually confined to observations up to distances of the order of 10 km. and used for the investigation of the rocks at comparatively small depths.

As the original disturbance is almost wholly outward, P movements in artificial earthquakes are much larger than S ones, contrary to what is found in natural earthquakes. S from explosions has hardly been convincingly recorded, though Rayleigh waves are produced, and their beginning looks rather like S.

The principle is that if two points are at distance x apart, z is the depth, v the velocity of a pulse at the greatest depth reached, and c the velocity at depth z, the time of transmission is

$$T = \frac{x}{v} + \int \left(\frac{1}{c^2} - \frac{1}{v^2} \right)^{\frac{1}{2}} |dz|,$$

accurate to quantities of the second order in the slope of any interface. The quantity under the integral sign has been called the *delay-depth coefficient* by A. W. Lee. If we denote it by η, and the depths of the first interface below the beginning and end of the path are H_1, H_2,

$$T = x/v + \eta(H_1 + H_2).$$

If several seismographs are arranged in a straight line, v and H_1 remain constant, but H_2 varies if the interface is not parallel to the outer surface. If two explosions

* Since this was written I have found that the Burton-on-Trent explosion of 1944 Nov. 27 gave a P, recorded at Puy de Dôme, and Sg was well recorded.

are made at distance l apart, at places where the depths are H_0, H_0', and the seismographs are placed along the straight line connecting them at distances $x_1, x_2 \ldots$ from the first, the time at x_n for the first explosion will be

$$T_n = \frac{x_n}{v} + \eta(H_0 + H_n)$$

and for the second

$$T_n' = \frac{l - x_n}{v} + \eta(H_0' + H_n).$$

The difference of these is independent of H_n, and if times are known for several values of x_n we can find v; then η is calculated, and the sums $T_n + T_n'$ vary only on account of the variation of H_n, which can therefore be found (Jeffreys, 1935 b).

A direct wave in the upper layer of sediments is sometimes unobservable over a long distance if that layer is chalk or clay; damping is strong unless the rock is hard or under considerable pressure. It is usual to start from a place where the stratum to be mapped either appears at the surface or has its depth found from a boring (and without such a check the geological interpretation will remain uncertain even if the form is known), but, given a starting point, the method can be used to follow the top of a given stratum for considerable distances.

Extensions are obviously possible to find the forms of several superposed layers.

The method has been much used in oil prospecting, especially in the United States and Persia, since the oil collects in the anticlines in a limestone or a salt dome. The highest point in the oil-bearing layer is located and provides the most promising place for a boring.

The most thorough published description of the method is by Bullard, Gaskell, Harland and Kerr Grant (1940), who used it for following the top of the Palaeozoic rocks through Eastern England; several Mesozoic strata were followed at the same time, and the results were checked by the existing borings. The agreement at these points was satisfactory, and the result is a substantial contribution to geology. Another application is to the investigation of the structure of the continental shelf; this was begun in the Eastern United States by M. Ewing (1937), and applied to the mouth of the English Channel by Bullard and Gaskell (1941). This work was unfortunately interrupted by the outbreak of war, but the preliminary results look extremely promising.

Reflexions have also been used, especially in California.

§ 15. ACKNOWLEDGMENT

By the kind permission of the Council of the Royal Astronomical Society, I have included in this *Report* a considerable part of the Report on Seismological Tables from the Annual Report of that Society for 1939.

GENERAL REFERENCES

Byerly, P., 1942. *Seismology* (New York: Prentice-Hall).
Gutenberg, B., 1932. *Handb. Geophys.*, **4** (Berlin: Borntraeger).
Heck, N. H., 1936. *Earthquakes* (Princeton: University Press).
Jeffreys, H., 1929. *The Earth* (Cambridge: University Press).
Jeffreys, H., 1935. *Earthquakes and Mountains* (London: Methuen).

OTHER REFERENCES

BASTINGS, L., 1935. *Proc. Roy. Soc.*, A, **149**, 88–103.

BULLARD, E. C. and GASKELL, T. F., 1941. *Proc. Roy. Soc.*, A, **177**, 476–99.

BULLARD, E. C., GASKELL, T. F., HARLAND, W. B. and KERR GRANT, C., 1940. *Phil Trans.* A, **239**, 29–94.

BULLEN, K. E., 1936. *Geophys. Suppl.*,* **3**, 395–401.

BULLEN, K. E., 1937 a. *Geophys. Suppl.*, **4**, 143–64.

BULLEN, K. E., 1937 b. *Trans. Roy. Soc. N.Z.*, **67**, 122–4.

BULLEN, K. E., 1938. *Brit. Assoc., Gray-Milne Trust.*

BULLEN, K. E., 1939 a. *Geophys. Suppl.*, **4**, 579–82.

BULLEN, K. E., 1939 b. *Geophys. Suppl.*, **4**, 583–93.

BULLEN, K. E., 1940 a. *Trans. Roy. Soc. N.Z.*, **70**, 137–9.

BULLEN, K. E., 1940 b. *Bull. Seism. Soc. Amer.*, **30**, 235–50.

BULLEN, K. E., 1941. *Trans. Roy. Soc. N.Z.*, **71**, 164–6.

BULLEN, K. E., 1942 a. *Bull. Seism. Soc. Amer.*, **32**, 19.

BULLEN, K. E., 1942 b. *Trans. Roy. Soc. N.Z.*, **72**, 141.

BYERLY, P., 1926. *Bull. Seism. Soc. Amer.*, **16**, 209–65.

BYERLY, P., 1930. *Gerlands Beitr.*, **26**, 27–33.

BYERLY, P., 1939. *Bull. Seism. Soc. Amer.*, **29**, 427–62.

BYERLY, P. and WILSON J. T., 1935. *Bull. Seism. Soc. Amer.*, **25**, 223.

CONRAD, V., 1925. *Mitt. ErdbKomm., Wien.*

CONRAD, V., 1928. *Gerlands Beitr.*, **20**, 240–77.

DAVISON, C., 1936. *Great Earthquakes* (Cambridge: University Press).

EWING, M., CRARY, A. P. and RUTHERFORD, H. M., 1937. *Bull. Geol. Soc. Amer.*, **48**, 753–802.

GUTENBERG, B., 1915. *Veröff. ZentBur. Int. Seism. Ass.*

GUTENBERG, B., 1914. *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.*, 125–76.

GUTENBERG, B., 1932. *Gerlands Beitr.*, **35**, 6–50.

GUTENBERG, B. and RICHTER, C. F., 1934. *Gerlands Beitr.*, **43**, 56.

GUTENBERG, B. and RICHTER, C. F., 1938. *Geophys. Suppl.*, **4**, 363–72.

GUTENBERG, B. and RICHTER, C. F., 1939. *Gerlands Beitr.*, **54**, 94–136.

JEFFREYS, H., 1926. *Geophys. Suppl.*, **1**, 371–83.

JEFFREYS, H., 1927. *Geophys. Suppl.*, **1**, 483–94.

JEFFREYS, H., 1928. *Geophys. Suppl.*, **1**, 500–21.

JEFFREYS, H., 1931 a. *Geophys. Suppl.*, **2**, 318–23.

JEFFREYS, H., 1931 b. *Geophys. Suppl.*, **2**, 329–48 and 399–406.

JEFFREYS, H., 1931 c. *Geophys. Suppl.*, **2**, 407–16.

JEFFREYS, H., 1935 a. *Geophys. Suppl.*, **3**, 310–43.

JEFFREYS, H., 1935 b. *Proc. Phys. Soc.*, **47**, 455–8.

JEFFREYS, H., 1936. *Bur. Centr. Séism., Trav. Sci.*, 14.

JEFFREYS, H., 1937 a. *Geophys. Suppl.*, **4**, 1–13.

JEFFREYS, H., 1937 b. *Geophys. Suppl.*, **4**, 50–61.

JEFFREYS, H., 1937 c. *Geophys. Suppl.*, **4**, 165–84.

JEFFREYS, H., 1937 d. *Geophys. Suppl.*, **4**, 196–225.

JEFFREYS, H., 1938 a. *Geophys. Suppl.*, **4**, 281–305.

JEFFREYS, H., 1938 b. *Phil. Trans.*, A, **237**, 231–71.

JEFFREYS, H., 1939 a. *Theory of Probability* (Oxford: Clarendon Press).

JEFFREYS, H., 1939 b. *Geophys. Suppl.*, **4**, 424–60.

JEFFREYS, H., 1939 c. *Geophys. Suppl.*, **4**, 498–533.

JEFFREYS, H., 1939 d. *Z. Geophys.*, **15**, 168–75.

JEFFREYS, H., 1939 e. *Geophys. Suppl.*, **4**, 548–61.

JEFFREYS, H., 1939 f. *Geophys. Suppl.*, **4**, 594–611.

JEFFREYS, H., 1940. *Bull. Seism. Soc. Amer.*, **30**, 225–334.

JEFFREYS, H., 1941. *Geophys. Suppl.*, **5**, 33–36.

JEFFREYS, H., 1942 a. *Proc. Camb. Phil. Soc.*, **39**, 48–51.

* *Monthly Notices of the Royal Astronomical Society, Geophysical Supplement.*

JEFFREYS, H., 1942 b. *Geol. Mag.*, **79**, 291–5.
JEFFREYS, H. and BULLEN, K. E., 1935. *Bur. Centr. Séism., Trav Sci.*, **11**.
JEFFREYS, H. and BULLEN, K. E., 1940. *Seismological Tables, Brit. Assoc., Gray-Milne Trust.*
LAMB, H., 1904. *Phil. Trans.*, A, **203**, 1–42.
LEHMANN, I., 1934. *Geod. Inst., Copenhagen, Medd.*, **5**.
LEHMANN, I., 1936. *Bur. Centr. Séism., Trav. Sci.*, **14**, 3–31.
LEHMANN, I. and PLETT, G., 1932. *Gerlands Beitr.*, **36**, 38–77.
OLDHAM, R. D., 1906. *Q.J. Geol. Soc.*, **62**, 456–75.
MACELWANE, J. B., 1930. *Gerlands Beitr.*, **28**, 165.
RICHTER, C. F., 1935. *Bull. Seism. Soc. Amer.*, **25**, 1–32.
SCHMERWITZ, G., 1938. *Z. Geophys.*, **14**, 351–90.
SCRASE, F. J., 1931. *Proc. Roy. Soc.*, A, **132**, 213–35.
SCRASE, F. J., 1933. *Phil. Trans.*, A, **231**, 207–34.
SIEBERG, A., 1932. *Handb. Geophys.* (ed. Gutenberg), **4**, 551.
STONELEY, R., 1928. *Geophys. Suppl.*, **1**, 527–32.
STONELEY, R., 1931. *Gerlands Beitr.*, **29**, 417–35.
STONELEY, R., 1935. *Geophys. Suppl.*, **3**, 262–71.
STONELEY, R., 1937. *Geophys. Suppl.*, **4**, 43–50.
STONELEY, R., 1939. *Geophys. Suppl.*, **4**, 461–8.
STONELEY, R. and TILLOTSON, E., 1928. *Geophys. Suppl.*, **1**, 521–27.
TURNER, H. H., 1922. *Geophys. Suppl.*, **1**, 1–13.
TURNER, H. H. 1926. *Geophys. Suppl.*, **1**, 425–46.
WADATI, K., 1928. *Geophys. Mag., Tokyo*, **1**, 162–202.
ZÖPPRITZ, K., 1907. *Nach. Ges. Wiss. Göttingen, Math.-Phys. Kl.*, **5**, 289–49.

X-RAY DIFFRACTION TECHNIQUES IN THE INDUSTRIAL LABORATORY

By H. P. ROOKSBY,
Research Laboratories of The General Electrical Company, Ltd., Wembley, England

§1. INTRODUCTION

The purposes for which x-ray diffraction methods have been utilized fall into two different, though related, categories. Firstly there is the wide field of investigation and analysis of crystal structures, and it is perhaps in this field that x-ray diffraction has had its most spectacular triumphs. Most of the early uses of x-ray diffraction were focused on the determination of atomic arrangements, and without this pioneer groundwork the more directly practical uses constituting the second category would probably not have been developed.

In this second category x-ray diffraction technique has become an analytical tool, supplementing with its own characteristic information results of investigations obtained by other physical and chemical methods. The possibility of using x-ray diffraction photographs for other purposes besides crystal-structure determination became apparent following the discovery of the Hull-Debye-Scherrer powder method, but practical industrial applications have also been found for some other methods originally developed for crystal-structure work. The use of the x-ray pattern of a crystalline substance as an identification mark or "crystal fingerprint" is obviously of practical value, but many other observations can be made by following certain alterations in appearance of the patterns. The evaluation of crystal size from the broadening of the x-ray reflections and the detection of lattice strain by observing the impairment of resolution of high-order reflections are but two examples.

This *Report* is concerned primarily with applications of industrial importance in the second category. Reference will not be made to actual crystal-structure determination except where it may be necessary for a proper understanding of a particular application. On some occasions, of course, a problem that in the first place may have seemed to permit of a relatively simple solution eventually involves the investigator in fundamental considerations of x-ray crystallography, but for the most part the actual evaluation of atomic arrangements is not required.

There has been a particularly rapid growth of interest in x-ray diffraction methods during recent years. In 1941 a symposium of papers was published in the *Journal of Scientific Instruments* on *X-ray Analysis in Industry*. Since that time several conferences have been held under the auspices of the Institute of Physics to discuss x-ray technique, and an x-ray Analysis Group of the Institute was formally constituted in 1943.* In America an x-ray and electron-diffraction section of the American Institute of Physics has been in existence for some years.

It is therefore not inappropriate that some account should be given in the

* See note in *Journal of Scientific Instruments*, 1943, p. 152.

Physical Society's *Reports* of the development and possibilities for practical use of this essentially physical tool in an industrial laboratory. It is perhaps in the large industrial laboratory, concerned with a variety of problems, that X-ray analysis can be employed to greatest effect, though there are many instances in which X-ray technique has been developed and applied to the control of a single industrial process.

It will be convenient to divide the article into two parts. In the first part some account is given of the use of the X-ray ionization spectrometer and of Laue patterns for the examination of orientation and texture of large single crystals. The second part is devoted to a discussion of the Hull-Debye-Scherrer powder method and its practical applications. A brief mention is made in this part of the varieties of X-ray tube which are utilized as sources of X radiation.

PART I. THE EXAMINATION OF LARGE SINGLE CRYSTALS BY X-RAY METHODS

§2. IONIZATION SPECTROMETER AND LAUE EXAMINATION OF SINGLE CRYSTALS

(i) *The X-ray ionization spectrometer*

One of the first methods of examining crystals by X rays employed a modified optical spectrometer. The crystal was placed on the table of the spectrometer,

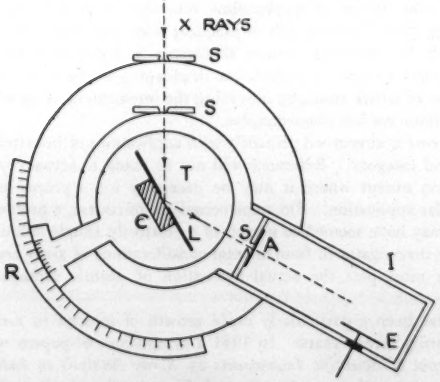


Figure 1.
Diagrammatic representation of principal features of X-ray ionization spectrometer.
S = Slits ; T = Turntable ; L = Locating plate ;
C = Crystal ; A = Aluminium window ; I = Ionization chamber ;
E = Insulated electrode to current-measuring circuit ;
R = Reading scale and vernier.

and the X-ray reflections were detected with a chamber of gas fitted in place of the analyser of the spectrometer. The gas was ionized when X rays entered

the chamber, and the degree of ionization was a measure of the intensity of the reflection. The arrangement is illustrated diagrammatically in figure 1.

The measurement of ionization current in the absence of modern electronic instruments was difficult, and, owing to the limitations on the size of the crystals which could be conveniently handled, after the pioneer work of the Braggs the method was not much used. Photographic methods, such as the rotating-crystal method, and later moving-film methods, found favour for crystal-structure investigations, in spite of the acknowledged difficulties of accurately determining intensities of reflections photographically.

Comparatively recently, the interest in the ionization spectrometer has been revived. Not only is the ionization method an ideal one for measuring intensities of reflections, but it has been found of practical value in other directions. An instrument designed primarily for structural investigations was described in 1936 by Wooster and Martin. Several novel features were introduced, one of them being the provision of means for automatically recording ionization current against the angular setting in the form of a graph.

More recently the ionization spectrometer has been simplified and employed in an industrial application. The particular problem that has been solved is the determination of crystallographic directions in large quartz crystals employed for the manufacture of piezoelectric oscillator and resonator plates. The measurement of the angle between a cut face and a particular atomic plane of a crystal is an obvious application for the spectrometer ; the angular setting of the atomic planes themselves, and not that of the exposed cut face, determines whether the x-ray wave-length is reflected or not.

(ii) *The determination of crystallographic directions in quartz*

Quartz crystallizes in the trigonal division of the hexagonal system, and may be referred to two principal axes. The optic or c axis corresponds to the height of the prism whilst the a axis corresponds to the side of the hexagonal base. In telecommunication work it is customary to refer to these axes as z and x axes respectively, and an x axis is also known as the electric axis. An obvious feature of the hexagonal structure is that there are three x directions intersecting at angles of 120° with one another.

The optic axis or z direction in quartz can be found to a fair accuracy by optical means, but the direction of the electric or x axis is difficult to locate in the absence of at least one well-developed and identifiable crystal face. In the past, faceted crystals were selected for cutting purposes, and these could be oriented, using classical crystallographic methods. Such a selection basis for the raw material became very uneconomic and wasteful as production demands increased, and some means had to be found for utilizing the large proportion of rough crystals that, although having no identifiable facets, were otherwise of good quality.

A very simple procedure is possible with the x-ray ionization spectrometer. In the first place the optical determination of the z axis can be checked to a high degree of accuracy. If a small section is cut perpendicular to the z axis the crystal is set on the spectrometer table so that the cut face contains the axis of rotation.

The angle of diffraction for the z plane may be calculated from known crystallographic data for quartz and the wave-length of the incident X rays. This enables the position of the ionization chamber to be fixed to receive the Bragg reflection when it occurs. The crystal table is then rotated until the maximum ionization current is observed. The difference, if any, in angular setting of the table from the Bragg angle for the z plane is a measure of the angle between the cut section and the true z plane. Two measurements, with the crystal turned between them through 90° about the z axis, will suffice to re-orient the z section accurately when an error is found.

Following the accurate location of the z section, etching or rough grinding methods may be employed to obtain a preliminary indication of the directions of the x axes in the z plane (Rivlin, 1940). It is then possible to cut a new face on the crystal that is approximately in the orientation required. This is examined on the X-ray spectrometer, using the settings appropriate for the particular crystal plane involved, and recut and ground until sufficiently precisely oriented.

American manufacturers have produced apparatus specially designed for this work.* The quartz-crystal X-ray goniometer manufactured by the Victor X-ray Corporation is shown in figure 2.† This instrument is simple to operate and has been robustly designed so that it may be used in factory control. The ionization chamber may be pre-set at the known angle corresponding with the crystal planes under investigation. There are a limited number of positions that will be required in practice. The degree of ionization is measured through a valve amplifier set at the front of the unit, the deflection of the needle of an ordinary milliammeter instrument being observed.

The crystal table is supported by a tapered bearing, and is attached to a long reading arm carrying the vernier of a steel divided arc on which angular positions are read. The arm is moved by hand over the relevant angular range until a kick of the meter needle occurs, and then the exact position found for maximum deflection in the region of the kick. Although there is no fine control on the adjustment, the finish of the instrument is so good that settings can be repeated within two or three minutes of arc.

The surface of the crystal under investigation is located against a vertical plate containing an aperture to allow the X-ray pencil to strike the crystal. The locating plate is fixed in position so that when it is parallel to the incident X-ray beam the vernier scale registers zero angle.

(iii) *Other applications of the ionization spectrometer*

The spectrometer or goniometer described above was specially designed for the quartz application. With the addition of a circular scale for the ionization chamber, to enable it to be quickly set in any predetermined angular position instead of at the few fixed points chosen for orienting quartz, the goniometer could be utilized equally well for other crystals. Two other industrially important crystals, synthetic sapphire and diamond, have also been subjects for crystallographic orientation studies, and would be suitably examined with the spectro-

* *X-ray Crystal Analysis Apparatus*, Philips Metalix Corporation, 419 Fourth Avenue New York ; *Quartz Crystal X-ray Goniometer*, General Electric X-ray Corporation, Chicago, U.S.A., and Victor X-ray Corporation, 15 Cavendish Place, London.

† Facing page 92.

meter. With crystals of small dimensions, like industrial diamonds, it is necessary to replace the vertical locating plate employed for large crystal surfaces by a crystal holder of appropriate design. This must enable the crystal to be rotated and moved in several directions to allow a chosen reference surface to be correctly oriented with respect to the axis of rotation of the spectrometer table and the direction of the primary x-ray pencil. A two-circle goniometer arrangement supports the crystal at the centre of two arcs that are at right angles to one another, and

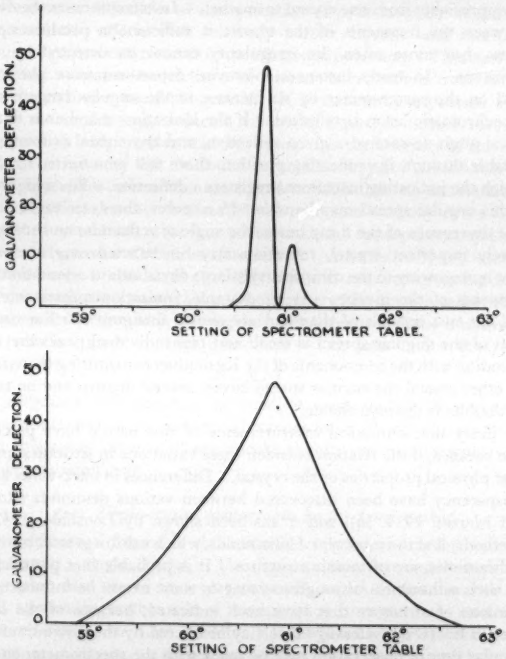


Figure 3. X-ray ionization spectrometer curves from diamond crystals.
Above: crystal with relatively perfect structure.
Below: crystal with disoriented mosaic structure.

ride on a pair of slides that can move in perpendicular directions. A more elaborate holder, that permits of a wide variety of movements of the crystal, has been described by N. Wooster (1943).

Besides orientation determinations, the x-ray ionization spectrometer may also be adapted for examining the geometrical perfection of structure of a crystal. Ideally the atomic planes in a single crystal extend without discontinuity to the external boundaries, but this ideal condition is probably never realized in practice.

There may be limited regions in the geometrically perfect condition, but these small fragments probably differ slightly in relative orientation. The crystal could thus be regarded as a mosaic, the fragments of the mosaic having extremely small angular inclinations to one another. It this were not true, it can be shown that, because of extinction effects (Darwin, 1914, 22), the intensities of x-ray reflections would be very much lower than those actually observed.

But the point of practical interest here is that the imperfections in structure do vary appreciably from one crystal to another. In extreme cases the disorientation between the fragments of the mosaic is sufficient to produce opacity or cloudiness, but more often the irregularity cannot be detected by ordinary optical means. In many instances, however, departures from perfection are indicated on the spectrometer by an increase in the angular range over which the monochromatic beam is reflected. If the ionization chamber is fixed at the theoretical angle to receive a given reflection, and the crystal is rotated on the crystal table through the reflecting position, there will be a finite angular range over which the indicating instrument registers a deflection. For a highly perfect crystal this angular spread may be about 15 minutes, the exact value depending upon the divergence of the x-ray beam, the angle of reflection, and so on. With a relatively imperfect crystal, reflection may be detected over an angle of a degree or more, owing to the comparatively large deviations in orientation between the fragments of the mosaic. As an example, ionization-spectrometer curves obtained for two crystals of diamond are shown in figure 3. For one crystal the value of the angular spread is small and two individual peaks are resolved, corresponding with the components of the Kα doublet constituting the x-radiation. For the other crystal the angular spread covers several degrees and no resolution of the α doublet is distinguishable.

It is likely that numerical measurements of this nature have practical significance because of the relation between these variations in structural perfection and other physical properties of the crystal. Differences in ultra-violet and infrared transparency have been discovered between various diamonds (Robertson, Fox and Martin, 1934, 36), and it has been shown by Lonsdale (1942), using x-ray methods, that the rarer type II diamonds, which exhibit greater transparency in the ultra-violet, are of mosaic structure. It is probable that physical characteristics such as hardness or toughness may to some extent be influenced by the imperfections of structure that have been indicated, because of the increased resistance to fracture by cleavage that may be offered by the discontinuities.

A similar type of observation may be made with the spectrometer on surfaces finished or treated in different ways. For instance, Kirkpatrick (1941) has examined x-ray reflections from cleavage surfaces of calcite. With a slightly ground surface detectable reflection occurred over an 11° range of angles of incidence, whereas reflection from an untreated cleavage surface was detected over only 2°. Kirkpatrick concluded that the large angular spread for the ground surface was attributable to the presence of disordered crystal particles; the surface layers had been to some extent broken up and disoriented by the grinding process. It is probable that the degree to which a crystal surface is broken-up and becomes disoriented varies with the method of preparation or finishing, and the spectrometer gives a useful measure of the differences.

Another method of approach is to compare the intensities of reflection from the crystal face. Peak readings on the indicating meter as the crystal face is moved through the reflecting position provide a simple means of comparison, although obviously there must be some form of calibration if an accurate relationship is to be established between the readings and the actual intensity of reflection. A roughly ground surface gives a much higher intensity of reflection than a surface which has been freshly cleaved or carefully etched to remove the disoriented fragmented surface layers.

There are some difficulties of interpretation of peak-intensity measurements owing to extinction effects with relatively perfect crystals, and more precise comparisons require measurements of the integrated reflection. This is an angular value derived by comparison of the total amount of reflections from the crystal face with the radiation incident upon it. It is commonly expressed as $E\omega/I$, where E is the total energy reflected when the crystal is turned through the reflecting position with angular velocity ω, and I is the intensity of the incident radiation. The quantity can be found in another way, by a relatively simple procedure, applicable when the spectrometer table has no automatic drive. With the ionization chamber fixed at 2θ, the reflection intensity curve is plotted by setting the crystal table at various positions in the appropriate angular range, and the area, A_R, under the curve determined. A measure of the intensity, I, of the direct beam is also made ; usually this will require a rotating sector to reduce the intensity to a value that can be registered by the ionization chamber. The integrated reflection $\delta\theta$ is the angular value which, with the intensity of the incident radiation, gives an area equivalent to the area under the reflection curve, i.e.,

$$\delta\theta = A_R/I.$$

The integrated reflection has the smallest value for ideally perfect single crystals, and increases as the structure becomes more imperfect, and disorientations and fragmentation into mosaic blocks slightly out of alignment with one another become significant.

(iv) *Single-crystal photographs by the Laue method*

A Laue diffraction pattern is obtained when a beam of polychromatic x rays is incident upon a single crystal. The Laue pattern is normally registered on a photographic film placed in a plane at right angles to the direction of the incident x rays on the further side of the specimen from the x-ray source (figure 4). Cameras similar to those used for the flat-film Debye-Scherrer technique are suitable.

The complete analysis of Laue patterns is in general complex, but the patterns have been utilized in industrial applications for orientation determinations. In this work, back-reflection patterns (figure 4), in which the film is interposed between the x-ray source and the specimen, have proved particularly useful, because large crystals can be handled without special preparation. The method has been developed by Greninger (1935 a, 35 b) for finding the orientation of cubic metal crystals. For cubic crystals the back-reflection Laue photographs are comparatively easy to interpret, the symmetry characteristics standing out well enough to be quickly recognized. Principal crystallographic planes give

rise to distinctive reflections that usually coincide with the meeting point of prominent rows or curved lines of Laue spots.

D. E. Thomas (1939) employed the back-reflection technique, but instead of the conventional flat-film arrangement the cassette was conical, with a vertical angle of 60° and axes coinciding with the direction of the primary x-ray pencil. The film had to be cut to a special shape to lie inside the conical cassette, and a method was worked out for making stereographic or gnomonic projections from the patterns and ascertaining the orientation of the crystals. By use of a conical cassette a large angular range of reflections can be recorded, and an improvement in the clarity of the patterns was claimed because reflections did not in general strike the film so obliquely as in the more conventional arrangement.

The back-reflection Laue method has also been employed for determining the orientation of synthetic sapphires (Brown, Chirnside, Dauncey and Rooksby, 1944) made by the Verneuil process. Examples of typical patterns are shown in figure 5.*

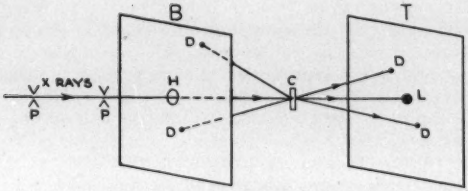


Figure 4. Diagram showing alternative positions for photographic film in relation to the crystal in Laue method.

C = Crystal ;
T = Film in position for transmission photograph ;
L = Lead stop for primary x-ray beam ;
B = Film in position for back reflection photograph ;
H = Hole punched in film ;
P = Pin-hole collimating system ;
D = Examples of Laue spots.

Another application of Laue patterns is to the study of the kind of imperfections in structure that have already been mentioned in connection with the ionization spectrometer. With a highly perfect crystal the size of the Laue spots is primarily a function of the size of aperture defining the x-ray pencil, the specimen-to-film distance and the thickness of the specimen. Moreover, the intensity distribution over each spot will be substantially uniform. But with a very imperfect crystal, disorientation between the elementary fragments will cause the reflection spots to spread out over large areas, and in some instances individual small "flecks" originating with the elementary units will be distinguished. The increase in size of the spots, and the discontinuities observed, thus enable an appraisement of the imperfections to be made. In figure 6 (*a*)* a Laue diffraction pattern of a crystal of diamond of very imperfect mosaic structure is shown. Back-reflection patterns can also be utilized, and one of an imperfect sapphire crystal is reproduced in figure 6 (*b*).*

The Laue method is further employed for the detection of certain types of

* Facing page 93.

strain or distortion, both in metal and in mineral crystals. The presence of this form of imperfection is revealed by a lengthening and tailing of the Laue spots in a radial direction, the phenomenon being generally known as radial asterism. Radial asterism may be observed for coarse-grain metal specimens that are photographed whilst under stress. It also appears on Laue patterns of metallurgical specimens in which there is permanent internal strain, such, for example, as may be present in a large rapidly cooled metal casting. The elongation of the Laue spots into radial streaks is due to changes in lattice orientation within the crystals, such disorientations being either associated with elastic strain or a more permanent deformation.

An example of radial asterism in Laue patterns of mineral crystals has been described by Hutton and Fankuchen (1938–40) in a study of the stilpnomelane group of minerals. They attribute the effect to imperfect packing of the structural units parallel to the layer plane. Rooksby (1945) has found similar effects in particles of a mineral impurity characteristic of many glass-making sands ; the x-ray pattern is shown in figure 6 (*c*).*

The information contributed by Laue patterns concerning the origin of non-Laue diffuse reflections and the connection with crystal dynamics has been discussed in detail by Lonsdale (1942–43).

PART II. THE EXAMINATION OF POLYCRYSTALLINE MATERIALS BY X-RAY METHODS

§3. THE X-RAY TECHNIQUE

Most industrially important materials are polycrystalline in nature, and the x-ray techniques that have been developed for the study of these materials are broadly described as powdered-crystal or powder methods. The description Hull-Debye-Scherrer method is also used, since the practicability of obtaining informative diffraction patterns from a powdered crystal was discovered almost simultaneously by Hull (1917 a–17 c) in America and Debye and Scherrer (1916, 17) in Germany.

The principle of the technique is well known. The substance, in the form of a finely divided powder, intercepts a collimated pencil of x rays of a single wavelength. The reflected rays travel along the generators of a family of coaxial cones, with common apex at the specimen. The angles of reflection are related by the Bragg equation $n\lambda = 2d \sin\theta$ to the various interplanar distances constituting the characteristic crystal structure of the substance. If the reflections are allowed to fall on a photographic film arranged perpendicular to the incident x-ray pencil a set of concentric rings is obtained. The film may be placed either on the side of the specimen remote from the source of x rays (the transmission method) or on the near side, when, before striking the specimen, the x rays pass through a hole punched in the centre of the film (the back-reflection technique).

Usually film-to-specimen distances are adjustable, and film cassette and specimen holders are located on some kind of camera track similar to an optical bench. The equipment, including collimator, specimen holder, and film

* Facing page 93.

cassette, manufactured by the Victor X-ray Corporation, is shown in figure 7. As previously mentioned, this equipment is also suitable for Laue photographs.

The dispositions of film cassette and specimen that have been indicated are those most commonly employed when the film is uncurved, though for special applications the film may be placed in oblique positions. For general analytical purposes, however, another form of camera is more appropriate. This is the cylindrical or circular powder camera, in which a narrow strip of film on the periphery almost surrounds the specimen mounted to lie along the axis. A diagrammatic illustration of the arrangement is shown in figure 8. In this arrangement the x-ray diffraction pattern takes the form of a series of slightly curved lines, and both low- and high-order reflections are recorded on the same film. References to powder or powdered-crystal photographs in accounts of x-ray work generally mean x-ray patterns taken with cameras of this type.

Before discussing important features of the technique of x-ray powder photography some brief reference will be made to generators of the x radiation itself.

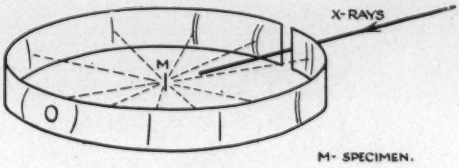


Figure 8.
Representation of formation of powdered crystal photograph in a cylindrical camera.

§4. SOURCES OF X-RADIATION FOR DIFFRACTION ANALYSIS

(i) *Demountable and sealed-off x-ray tubes*

X-ray tubes for crystal diffraction work fall into two main classes. Demountable tubes have been widely utilized in Britain, but completely sealed-off vacuum tubes have been more generally favoured in America and on the continent of Europe. The hot-filament electron tube of both types is most common, although a few laboratories are using the Müller type of gas tube.

There are certain advantages in employing a demountable electron tube both in respect of economy and flexibility. It is convenient (and often essential) to have available a number of different x-ray wave-lengths, and these may be provided by having a range of x-ray targets tipped with different metals. With a demountable tube it can be arranged that the targets are interchangeable with little difficulty, whereas several complete sealed-off x-ray tubes would be needed to provide the same flexibility. The target face can be cleaned at intervals to get rid of tungsten deposited from the filament and so the quality of the x radiation maintained at high level. Moreover the tungsten filament itself can be renewed as required and the life of the x-ray tube extended almost indefinitely. Because the components can be renewed or renovated at intervals, the tube may also be

Figure 2. Photograph of x-ray spectrometer made by the Victor X-ray Corporation for the determination of orientation of quartz crystals.

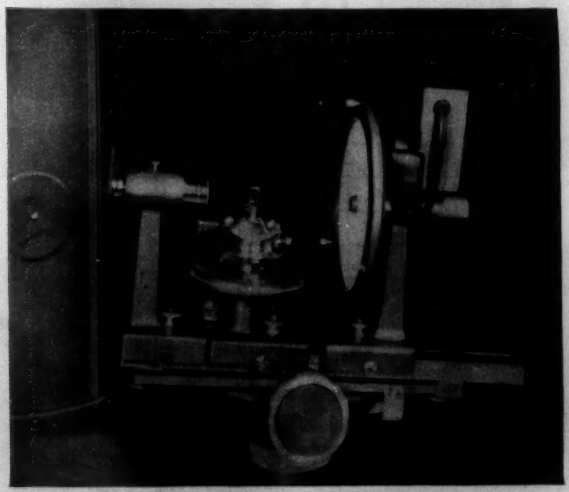
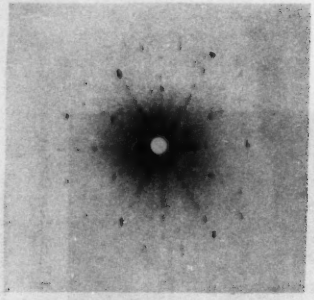


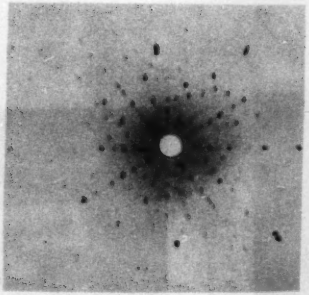
Figure 7. Flat film cassette, crystal holder and collimator of Victor X-ray Corporation's x-ray diffraction equipment.

(*Reproduced through courtesy of Victor X-ray Corporation, London.*)

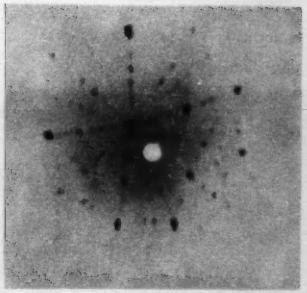
To face page 92



Quartz.
X rays normal to (001) plane.

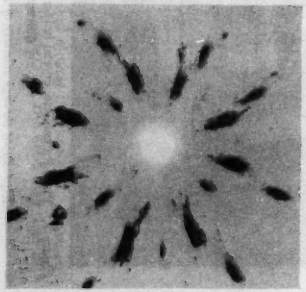


Sapphire.
X rays approx. normal to (110) plane.

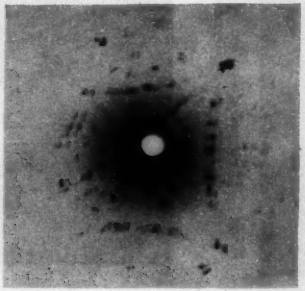


Tungsten.
X rays incident in oblique direction.

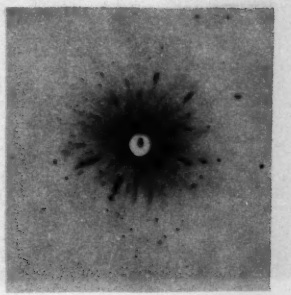
Figure 5. X-ray Laue back-reflection patterns (unfiltered radiation, distance 2·5 cm.)



Diamond.



Sapphire.



Particles of a mineral from a glass-making sand.

Figure 6. Laue patterns of imperfect crystals.

operated at substantially higher loads than is permissible with a sealed-off tube of similar design ; the rating of the latter must be relatively conservative if a reasonable life is to be obtained.

On the other hand very efficient pumping equipment must be an integral part of such an X-ray set, because with a demountable tube continuous pumping is essential. This means that in general more time must be expended on maintenance and operation of the X-ray apparatus itself than when a set with a sealed-off tube is employed. There is little doubt that for single industrial applications or routine control, when flexibility is of little importance, an X-ray set with a sealed-off tube has much to recommend it.

For some applications it is useful to have a transportable equipment that can be moved to the place at which the specimen is located. This may be desirable when X rays are being utilized for the determination of stress in large metallurgical pieces. A transportable set housing a sealed-off tube has been manufactured by the G.E. of America.* The Dexrae unit manufactured by Messrs. Adam Hilger, London, is also mounted on a movable trolley.†

(ii) *Beryllium windows*

An important discovery of recent years is how to prepare beryllium discs and sheet in a form suitable for use as windows of X-ray tubes. It was found by Claussen and Skehan (1942) that additions of 0·25 to 0·5% of titanium or

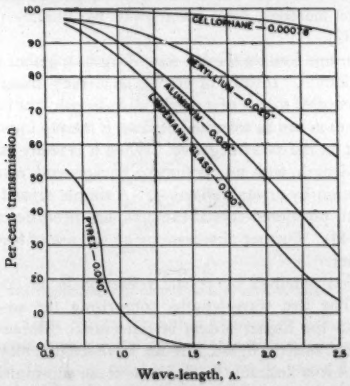


Figure 9. X-ray transmission of window in X-ray tubes for diffraction.
(*Reproduced, by kind permission, from the 'Review of Scientific Instruments'*.)

zirconium to beryllium eliminated the tendency to porosity and embrittlement and resulted in relatively malleable products that could be hot-rolled. Discs

* See note in *Journal of Scientific Instruments*, 1943, p. 51.
† Adam Hilger publication No. 257/3.

can be cut from vacuum cast ingots, and if further reduction in thickness is required the discs are hot-rolled.

Owing to its low atomic weight, beryllium in the window of an X-ray tube will absorb a much smaller proportion of the radiation from the target than aluminium of equivalent thickness; this is particularly so with the longer wave-lengths from iron and chromium. Figure 9 shows curves due to Brackney and Atlee (1943) comparing the X-ray transmissions of different possible window materials in the wave-length region between 0·5 and 2·5 A. From these it can be seen that there is a substantial gain in transparency of a 0·02″ section of beryllium over a foil of aluminium of thickness 0·001″, particularly with wave-lengths longer than 1·5 A.

(iii) *Range of useful X-ray wave-lengths*

The elements in the periodic table between copper and chromium give a range of wave-lengths suitable for the majority of applications, though a molybdenum target or a silver target may be desirable for Laue photographs. In practice most requirements of the powdered-crystal technique will be met by the respective radiations from copper, cobalt and chromium. To obtain monochromatic radiation it is necessary to filter the component from the characteristic K radiation by preferential absorption in an appropriate metal foil. The metal of the foil is chosen so that its K absorption edge lies between the α and β wave-lengths, and the various thicknesses and metals required with a wide range of target materials have been given by Bradley (*v.* Edwards and Lipson, 1941, p. 132).

Filtered X-ray beams contain the Kα wave-lengths together with a proportion of continuous X radiation. Ideally, of course, as already mentioned, the Debye-Scherrer method requires X rays of a single wave-length, but for many purposes the continuous radiation can be tolerated because it merely imposes an additional general background on the X-ray negative. When it is really essential to employ X rays of one wave-length, with no continuous background, resort must be made to monochromatization by crystal reflection. A simple crystal monochromator, suitable for use with powdered-crystal cameras, has been described by Donnay and Fankuchen (1944). Further observations on the use of focusing monochromators are made later.

The Kα radiation obtained by crystal reflection is, of course, not strictly monochromatic. The two wave-lengths comprising the so-called α doublet become separated in the higher orders of reflection. Moreover wave-lengths equivalent to one-half, one-third, etc., the Kα wave-lengths may also be reflected by the crystal, but if it is desirable to eliminate these sub-multiple wave-lengths this can be done by limiting the X-ray tube voltage below the potential required for their excitation.

(iv) *Rotating-target X-ray tubes*

The electrical input to an X-ray tube, and consequently the intensity of the emergent X-ray beam, is limited by the rate at which heat energy can be dissipated from the focal area on the target face. If the focus of the electron stream is located towards the edge of a moving target of large diameter the heating is distributed over a large area, so that much higher loading is per-

missible. The optimum conditions for stationary and moving targets have been studied theoretically by several investigators, including Müller (1931), DuMond, Watson and Hicks (1935), and de Graaf and Oosterkamp (1938).

Early work on rotating-target X-ray tubes for radiography was done by Bouwers (1930). One of the first moving-target tubes specially designed for X-ray diffraction work was described by Clay (1934), and later Müller and Clay (1939) built a 50-kilowatt generator for use in the Royal Institution, London. One of the chief problems of design is that of bringing out a hollow shaft, on which the target is mounted, through a vacuum-tight joint in which it may be spun. Cooling water has to be carried through the shaft to, and away from, the back of the target face. Müller and Clay (1939) used a vacuum stuffing-box containing a three-stage hat-leather and rubber-ring arrangement.

Astbury and Preston (1934) suggested a different arrangement, using the Torricellian vacuum principle. With the axis of the X-ray tube vertical, a long

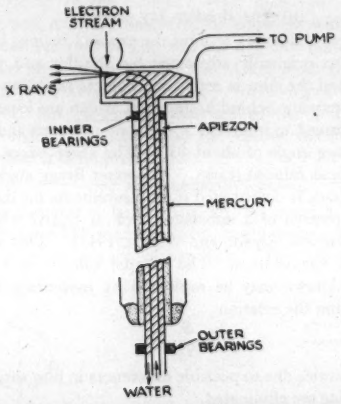


Figure 10.
Diagram showing principle of mercury-sealed rotating-target X-ray tube.

shaft carrying the target passes through a barometric column of mercury. In order to prevent mercury vapour getting into the X-ray tube and making its operation unstable, the inner mercury meniscus is covered with Apiezon low vapour-pressure oil. The diagram of figure 10 shows the basic features of the arrangement. The development of this principle and the construction and operation of an X-ray tube designed accordingly have been described by MacArthur (1944).

Rotating-target X-ray tubes would be advantageous for many applications in X-ray diffraction work. Not only would the reduction of the relatively long photographic-exposure times at present necessary be in itself very useful, but observations that are now difficult or precluded with stationary-target tubes would become practicable.

§5. DESIGN OF CYLINDRICAL POWDER CAMERAS

(i) *Requirements*

Although in the early days of x-ray powder photography cameras of very small diameter (4 or 5 cm.) were employed, the need today is for larger diameters to give a greater separation of the reflections. In a number of laboratories in Britain cameras 9 or 10 cm. in diameter are employed, and a still greater diameter of 19 cm. is also in use.

Modern cylindrical powder cameras must be carefully designed to ensure that patterns of high definition and clarity are obtained. The basic principles to be followed in the construction of a camera of 9 cm. in diameter have recently been examined by the X-ray Analysis Group of the Institute of Physics, and the final recommendations have been published.* An accurately located specimen holder, precisely designed collimator from which no extraneous scatter is possible, beam trap, and knife edges to provide reference shadows on the exposed film, are the chief requirements in a practical powder camera.

(ii) *The Bradley-Jay camera*

The use of knife-edge shadows to limit the exposed regions of the powdered crystal photograph was originally advocated by Bradley and Jay (1932). In the Bradley-Jay method the film is set in the camera as shown in figure 11 (*a*), the ends of the film passing behind knife edges which are located on the cylindrical surface and arranged to intercept x radiation reflected and scattered from the specimen at a Bragg angle of about 85°. The sharp edges to the exposed sections of the film act as fiducial marks. The exact Bragg angle corresponding with the fiducial shadows is determined by measurements on the camera, or by calibration with the pattern of a substance such as quartz whose lattice constants are accurately known (Lipson and Wilson, 1941). This angle θ_k is then used as a constant of the camera. The angular value θ of a reflection on a photograph of an unknown may be evaluated by measuring linear distances along the film and using the relation

$$\theta = \theta_k \frac{S}{S_k} .$$

By this means uncertainties due to possible differences in film shrinkage following photographic processing are eliminated.

Bradley and Jay (1932) also showed that in the determination of lattice constants various systematic errors, due, for example, to the divergence of the x-ray beam, eccentricity of centreing of the specimen, absorption in the specimen, etc., could be compensated by plotting an extrapolation curve. The majority of the errors vanish when the Bragg angle $\theta = 90°$, and in the case of cubic crystals, for instance, the lattice parameters calculated for the different reflections on the powder photograph show a systematic variation with θ. The best function of θ against which to plot the calculated values depends to some extent upon the magnitude of the contributions from the different sources of error. The problem has recently been discussed in detail by Taylor and Sinclair (1945). Although Bradley and Jay (1932) originally advocated $\cos^2\theta$ because eccentricity errors are linear with this function, Buerger (1942) suggests that, with accurate camera construction, absorption is the source of greatest error, and a $\cot\theta \cos^2\theta$ plot

* *Journal of Scientific Instruments*, 22, 57, 1945.

gives the closest approach to a straight line. Nelson and Riley (1945) prefer the function $\frac{\cos^2\theta}{\sin\theta} + \frac{\cos^2\theta}{\theta}$, which they show experimentally to give practically a straight line over the whole range of angles.

With crystals of hexagonal and tetragonal structure two lattice parameters are involved, but extrapolation methods are applicable, and the procedure required is summarized by Lipson and Wilson (1941). With more complex structures the difficulties of indexing the reflections become considerable and lattice-parameter determinations from powder photographs are in general impracticable.

(iii) *The van-Arkel camera*

An alternative method of setting the film in the camera is that suggested by van Arkel (1926), and illustrated diagrammatically in figure 11 (*b*). If D is the linear distance between corresponding lines of Bragg angle θ, $\theta = \frac{\pi}{2} - \frac{D}{4R}$, where R is the radius of the camera. With lines near Bragg angle 90° D is relatively small, and it is not necessary to know R with high accuracy. In practice R is measured to 0·01 cm. and the angle of the camera θ_k utilized as in the Bradley-Jay arrangement. The van Arkel method of setting the film is favoured by many workers, especially for metallurgical applications (Lipson and Wilson, 1941), but for general analytical problems it is preferable to have the strong low-order reflections near the centre of the negative, and the Bradley-Jay method is to be preferred.

(iv) *Large-diameter cameras*

With powder cameras of larger diameter, for example 19 cm., it is usual to divide the film into two halves (figure 11 (*c*)) as otherwise a single exposed negative is rather long for convenience of handling. The exposed portions of each

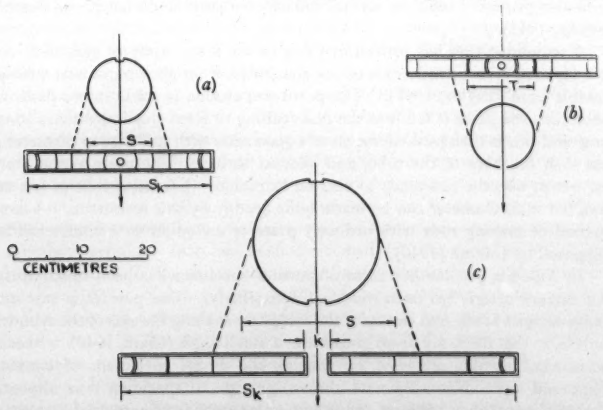


Figure 11. Diagram (after Lipson) showing usual film arrangements in cylindrical powder cameras and the types of film resulting from each.

half are limited by fiducial shadows, and two angles of the camera corresponding with the shadows are then required. To simplify measurements and calculations of θ values it is usual to employ a fixed linear distance, k, the distance between the low-angle knife edges, instead of the angular value. The linear distance between a given pair of lines is obtained by adding k to the sum of their distances along the films from the respective fiducial edges. Owing to differences in film shrinkage, this procedure may introduce an error because of an uncertainty as to the exact value of k, but the error is so small that it can be neglected. Details of construction of a 19-cm. diameter camera, that has found applications in several different fields, have been given by Bradley, Lipson and Petch (1941).

(v) *Mounting the specimen*

Several different methods are in common use for mounting the powder specimen along the axis of the camera. The most usual technique is to mix the powder with an adhesive such as Canada balsam, starch paste, or gum tragacanth, and then to apply the mixture to a hair or fine glass fibre, finally rolling out into a thin cylinder about 0·3 mm. in diameter. The hair, after attaching to the specimen holder, is kept taut with a small lead weight, and if the plane of the camera is accurately horizontal, may be quickly aligned with the axis of the camera. The glass fibre is stiff and does not require a weight, but it does not, of course, automatically lie parallel to the axis of the camera, and is, therefore, a little more difficult to "centre" than the hair.

Another method of mounting the powder is to fill thin-walled glass capillary tubes about $\frac{1}{2}$ mm. in diameter. Lindemann-type glass (Mazeleo, 1940), containing only lithia, beryllia and boric oxide, is very transparent to the X radiation normally employed, and is suitable for the capillary tubes. Pure silica tubes have also been used and are essential for high-temperature work. Methods for making tubes of cellulose acetate and other organic media have been described (Fricke, 1941).

A technique that has proved of value in the X-ray study of graphites, cokes and carbons is to extrude rods of the specimens from glass barometer tubing of suitable bore (Taylor, 1941). The powdered carbon is made into a paste with Seccotine, the paste is fed into the glass tubing to form a cylinder some 10 mm. long, and this is then pushed out on to a glass slide with stiff wire of diameter just less than the bore of the tube, and allowed to dry. The paste should not be too wet or else the rod tends to curl on extrusion. Extruded rods of not more than 0·4 mm. diameter can be made quite readily by this technique. A similar method of making rods with ordinary paste or collodion as a binder has been suggested by Lukesh (1940).

In America yet another form of powder specimen has been much utilized, and camera design has been modified accordingly. The powder is cast into a wedge-shaped block, and the tip of the wedge is set along the axis of the cylindrical camera, so that the X-ray beam grazes it at a small angle (Clark, 1940). Specially constructed moulds are used for making the wedge specimens of consistent shape and size. Naturally with the wedge type of specimen it is impossible to obtain a complete series of reflections on either side of zero angle; only one quadrant or semi-circle is actually completely recorded, although a few of the

low-order reflections in the other quadrant, probably of reduced intensity, will also be obtained. However, for many straightforward identification problems it may be quite sufficient to have only one section of the diffraction pattern registered.

To return to the form of specimen and camera customarily employed in Britain, it is usual to mix a "diluent" with materials containing elements of high atomic weight. The diluent is necessary in order that the mounted specimen shall be moderately transparent to the x radiation, as otherwise the balance is upset between the densities of low- and high-order reflections. A suitable medium for a diluent is powdered gum tragacanth ; as many as seven or eight parts by volume are required, for example, for a substance such as tungsten oxide.

The holder which supports the cylindrical-shaped specimens, fibre, thread, capillary tube, or extruded rod is arranged so that it can be continuously rotated during the photographic exposure. It may, for example, be carried in a ball-race and coupling provided with a small synchronous motor outside the camera cover. Rotation is desirable to ensure that the random orientation condition is satisfied ; it increases the number of crystal facets in the powder particles that will take part in any given reflection. With the wedge-shaped form of specimen provision is usually made for oscillating the wedge.

It is usual to arrange for the cover of the camera to form with the base an evacuable enclosure. To displace the air in the camera either by evacuation or by filling with hydrogen is a useful refinement of technique. It reduces the background scatter that occurs with air and improves the clarity of the photograph, particularly in the low-order regions.

Although the conventional techniques described above generally give the most consistent results, it is possible to obtain satisfactory powdered-crystal photographs from other forms of specimen and in many other ways. It is sometimes convenient, for example, to employ a flat section or strip, whose surface may be inclined at an angle of a few degrees towards the incident x-ray beam. The strip may be held by means of the conventional specimen holder and adjusted so that the axis of the camera lies in the surface to be investigated. Provided the width of the strip is limited to about 2 mm., no adjustment of the slit system of the camera is usually found to be necessary to obtain adequate definition of the reflections. The *glancing-angle* reflection method is very useful when surface films are involved, and if mechanism is provided for giving the specimen a small angular reciprocating motion during the exposure, the scope of the technique is still further increased.

Another technique that has been developed for the examination of clay and soil minerals is known as the *ordered-aggregate* method (Hendricks and Alexander, 1939 ; Nagelschmidt, 1941). A large number of the minerals found in clays and soils have a layer-lattice structure, and one of the means of identification is by the determination of the basal-plane spacings, which are characteristic for each clay substance. If a suspension of one of these clay minerals is allowed to settle out slowly, the particles orient themselves parallel to the basal plane. Specimens are prepared by allowing a small amount of a suspension to evaporate to dryness in a ground-glass ring attached to a cover slip

with petroleum jelly. The "ordered aggregate" thus obtained may be mounted on a glass fibre, and used either in the conventional powder camera or in a camera specially designed for the work. The plane of the specimen flake is inclined to the x-ray beam at a glancing angle of 2 or 3° or, better, the specimen is oscillated through a few degrees, when basal-plane reflections in several orders occur. The camera should be capable of registering reflections with interplanar spacings up to about 20 A.

§6. FOCUSING CAMERAS

Focusing cameras, though hitherto less widely used than the conventional Debye-Scherrer circular camera, have some features that may lead to their adoption for certain applications. The design follows the focusing property of a circle, the x-ray beam, specimen, and film being so located that the diffracted rays come to a focus at the film. In the symmetrical back-reflection focusing camera, a typical example of which is described by Owen and Pickup (1932), the slit, powder specimen, and film all lie on the circumference of a cylindrical camera. The powder specimen is placed on the curve of the camera directly opposite the slit, from which a divergent x-ray beam emerges, as shown in the upper diagram of figure 12. The photographic film with a hole punched in the centre is set in the back-reflection position, as in the van Arkel setting for the ordinary powder camera.

A considerable area of specimen is covered by the divergent x-ray beam, so that the photographic-exposure times are short, and, in consequence of the geometrical properties of the circle, all reflections are sharply focused at the film. The limitations are that low-order lines are difficult to record satisfactorily and, except for Bragg angles near 90°, the reflections strike the film at oblique angles of incidence.

This focusing principle is applicable in the design of back-reflection cameras with flat-film cassettes. If the entrance diaphragm is placed some little distance in front of the cassette itself, a comparatively divergent beam can be employed to effect short photographic exposures. But, provided the film to specimen distance is adjusted appropriately, a high-order reflection will be brought to a focus approximately at the film. The arrangement is, of course, far from perfect; the flat specimen usually employed will not focus the reflections exactly.

Practical value attaches to focusing cameras because they can be employed particularly usefully with a bent crystal monochromator. Bozorth and Haworth (1938) have described the preparation of bent crystals of rock salt, for example, for use as monochromators. The arrangement is shown in figure 12(*a*). The crystal is bent to a radius $2R$ and the reflecting face cut to a radius R. The source of x rays, X, the crystal, and the slit of the camera are arranged on the circumference of the circle of radius R, and only x rays of the selected wave-length will be focused on the slit. With this arrangement the whole energy of the monochromatized x-ray beam may be utilized, whereas much energy is wasted in a flat crystal monochromator, such as that appropriate for the ordinary Debye-Scherrer camera.

The bent crystal monochromator can also be employed with a focusing camera for low-order reflections. The radius of the camera and the reflecting

crystal must be chosen to allow the convergent x-ray beam to come to a focus on the film set in the front-reflection position, as shown in figure 12(*b*). The film and specimen positions are thus reversed as compared with the symmetrical back-reflection camera, and the convergent x-ray beam must be transmitted by the powder specimen to enter the camera.

Guinier (1939) favours an elastically bent quartz plate instead of the plastically deformed rock-salt crystal. Special holders are required to keep the quartz plate bent to the appropriate angle.

The advantages of these arrangements are that the crystal monochromatized x radiation can be utilized without long photographic exposures. This has potentialities in industrial applications, as well as in such fundamental problems as diffuse x-ray scattering by thermal agitation. The use of monochromatized radiation is naturally accompanied by a reduction of background on the negatives, because of the elimination of the white-radiation scattering. This increases

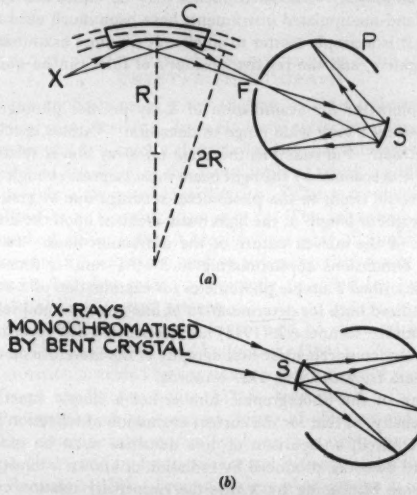


Figure 12. Diagrammatic representation of arrangements for use of focusing cameras with radiation monochromatized by a bent crystal.

X = X-ray source. C = Bent crystal.
S = Specimen. P = Photographic film.
F = Focus of incident monochromatic x-ray beam.

the sensitivity to detection of faint reflections and thus assists quantitative analysis at low concentrations. Photometry for the comparison of line densities is also simplified by the reduction of background, and the details of the crystal diffraction pattern proper more readily interpreted.

Monochromatized radiation is also required for the study of scattering at low angles. Warren (1936), Kratky (1942), and Biscoe and Warren (1942)

have shown that the form and distribution of scattering in the low-angle region rom, say, 0° to 1°, may be employed to provide information about particle size and shape. Riley (1943) has utilized the technique for studying structural differences between coals of different rank. Guinier (1943) has shown how the focusing principle with bent-crystal monochromator can be applied to the problem. The camera for low-angle scattering study must be evacuated, or at least hydrogen-filled, during the photographic exposure, so that no ambiguity arises because scattering from air is superimposed on the photographs.

§7. MICROPHOTOMETERS FOR X-RAY DIFFRACTION PATTERNS

Although comparison of line densities can be made with a fair degree of accuracy by visual inspection, for more accurate quantitative work it is necessary to employ photometric methods. Microphotometers for scanning x-ray negatives are generally based upon the use of the photocell, though in some instruments a thermopile is employed. The instruments may be either self-recording or hand-driven. Hand-manipulated instruments have been much used for powder photographs, for it is a simple matter to single out for close examination special details on the negative, and the relative positions of lines can be quickly ascertained.

Any microphotometer for examination of x-ray powder photographs must be capable of covering a very wide range of densities. Various special requirements have to be met. For example, the grain on x-ray film is relatively large, so that the area of film scanned by the light beam must be great enough to suppress fluctuations that could occur in the photometer readings due to grain. On the other hand, the height or length of the light beam incident upon the film must not be great, because of the curved nature of the diffraction lines. In practice a light image with dimensions approximating to $3 \times 0{\cdot}1$ mm. is found suitable. Jay (1941 a) has described a simple photometer for examination of x-ray patterns which has been utilized both for determination of line positions and for measurement of line densities. Ronnebeck (1943) has described a self-recording instrument, and gives a very comprehensive bibliography of the literature on the subject of microphotometers from the year 1927 onwards.

The blackening of the photographic film is not a simple function of the incident x-ray intensity, so that for the correct evaluation of reflection intensities, when these are required, comparison of line densities must be made against a series of standard densities produced by radiation of known intensity. In the case of photographic blackening by x rays the reciprocity relation connecting time and intensity holds to a high degree of accuracy. The film may therefore be calibrated by exposure to the primary beam for times varying by equal increments, when a series of densities produced by x-ray intensities differing by known increments is obtained. The calibration is most conveniently done with the aid of an appropriately designed rotating sector, such as that described by Robertson (1943). The sector is so cut that it produces what is called a step wedge, a series of steps of varying density representing equal increments of x-ray intensity. For precision work, photographic processing of negatives has to be carefully standardized, or else the step wedge must be imprinted for comparison purposes on every film.

Besides the conventional form of microphotometer, various devices have been suggested that would show the x-ray powder diffraction lines as peaks of different heights on a graphical trace. Tatel and Hultgren (1938) for example employed a photocell scanning device and oscillograph, so that the lines appeared as peaks on a curved trace on a cathode-ray tube screen. This was mainly intended for determining line positions accurately, the peaks being centred on a fiducial line on the screen by moving the film carriage across the light beam.

A recent proposal by Davey, Smith and Harding (1944) does not perhaps constitute a microphotometer in the usual sense, but is an arrangement whereby the powder photograph is represented in the form of a graphical trace. The apparatus records the x-ray pattern directly as an intensity-against-angle graph. The intensities of diffracted beams are measured by means of a Geiger-Müller counter mounted on a slowly-moving spectrometer arm. The incident radiation is Mo K and, by using balanced filters of Zr and Sr and appropriate electrical devices, the pattern is essentially that which would have been obtained with a pure Mo Kα beam.

§8. CHEMICAL IDENTIFICATION WITH THE X-RAY POWDERED-CRYSTAL PHOTOGRAPH

One of the applications of x-ray analysis of chief industrial importance is that of chemical identification. In so far as the crystal structure of a substance or chemical compound is a unique characteristic of that substance, the distribution and intensities of the lines of the x-ray pattern, which are determined by the structure, provide a means of identification. The x-ray pattern has been very aptly described as a "crystal fingerprint". The important contribution that the x-ray method makes is that it identifies the state of combination of the constituent elements of a substance, and so covers another aspect of analysis, using the term in its broadest sense, not directly provided for by the quantitative determination of the elements present.

A number of different examples in the field of chemical identification have been described by Bunn (1941) and by Rooksby (1941). As an illustration of the kind of information provided by a systematic x-ray study of a substance prepared under different conditions, figure 13* is reproduced. In this figure x-ray powder photographs of the various forms of alumina and its hydrates are shown. The alumina-water system has been thoroughly investigated by Weiser and Milligan (1934) and by Laubengayer and Weisz (1943) following preliminary studies by Rooksby (1929). From the x-ray photographs it could be concluded that two polymorphic crystal forms of the hydroxide can be made, that there are two forms of the monohydrate, and that a gamma and an alpha form of anhydrous alumina exist. Gamma alumina is formed as a metastable phase when the hydroxides or the monohydrate böhmite are decomposed below 1000° c. Transformation to alpha alumina is effected above 1000° c., but alpha alumina is produced directly by decomposition of the monohydrate diaspore. The broadened reflections on the gamma-alumina photograph are indicative of very small ultimate crystal size. It is this feature of γ alumina that is in part responsible for its high adsorptive capacity and its value as the essential ingredient in "activated alumina" driers. Line-broadening is discussed in detail later.

* Facing page 108.

It will be obvious that two or more crystal modifications of a substance are invariably distinguished by different x-ray patterns. The distinction by x-ray methods is often of considerable practical importance. This is the case with silica firebricks. The raw material is quartz, and it is necessary to burn the bricks during manufacture at 1300 to 1400° c. to convert the quartz to cristobalite and, or, tridymite. The efficacy of the firing treatment may be tested by x-ray examination of an appropriately selected specimen.

With silica firebricks, other physical methods of testing, e.g. thermal expansion or microscopic inspection, provide alternatives to x-ray analysis, but there are instances in which only the x-ray technique is practicable. For example, the examination of titania pigments for the two crystal forms rutile and anatase can only be carried out by x-ray means. This is again an example in which a heat treatment is employed during manufacture to convert one crystal form to another, in this instance anatase into rutile. Rutile pigments are now favoured because of their higher refractive index and greater covering or hiding power, and x-ray examination has proved of great assistance in the establishment of satisfactory methods of preparation. As little as 1% of residual anatase can be detected in a titania pigment using the x-ray powder photograph.

X-ray analysis is employed appropriately for the identification of crystalline constituents which are embedded in a non-crystalline matrix. Examples of this kind of application are the identification of the substances responsible for the opacity of opal glasses (Ryde and Yates, 1926; Rooksby, 1939), the identification of de-vitrification products in glasses, and the determination of the nature of the substances formed by reaction between glass and the refractories on which it is melted. In the last instance crystalline matter in the form of "stones" may be carried into the glassware being manufactured, and x-ray technique has proved very efficacious for the identification of these stones.

Another direction in which the x-ray method is of unquestionable value lies in the analysis of very small amounts of material. The specimen from which the x-ray diffraction pattern is obtained uses a minute quantity of the substance under investigation. When a large amount of material is available this does in fact mean that special care is necessary in sampling to ensure that the specimen chosen is representative. But it often happens that identification is required of some material of which only a very small quantity is available, and in these circumstances x-ray technique is employed to advantage. In experimental development work on vacuum devices of various kinds, for example, foreign deposits may occur, and identification is required so that the source may be traced. X-ray methods have also been employed for the examination of oxide films and other products of corrosive action on metals, for the identification of inorganic minerals in lung tissues in connection with research on silicosis (Nagelschmidt and King, 1941; Frankel, 1944), for the analysis of débris formed in instrument sapphire jewel cups, for the determination of the nature of small residues in semi-microchemical analyses, and so on. In some examples adherent films may be photographed by surface-reflection technique with the usual cylindrical powder camera, an appropriately dimensioned specimen being chosen. The value of this technique has been stressed by Imhoff and Burkardt (1943). In an examination of boiler scale they were able to reveal the non-homogeneous character of flakes of scale by means of separate photographs from the two surfaces.

Sometimes it may be more convenient to remove the film in question as a powder in order to obtain an X-ray pattern, and it is useful to have a diamond available for this purpose. Diamond is so much harder than most other materials that films are invariably removed very easily, and the "scrapings" are collected to provide the required specimen. The various thin interlayers that develop on refractory materials during use in glass-melting tanks have been examined in this way. There are many thin films and layers, which at first might be thought to be more appropriately studied by electron diffraction methods, that have been found to be suitable subjects for analysis by X-ray methods.

Some estimate of the relative proportions of two or more phases present in a mixture can be made from the relative densities of lines characteristic of the separate phases. In particular applications, sensitivity can be increased by photometric comparison of line densities. Usually a series of photographs of synthetic mixtures will be required, as in quantitative spectrographic work, to provide standards of comparison. Quantitative determinations have also been made employing the internal-standard technique, in which a definite proportion of some selected crystalline substance, to be used as standard, is added to the unknown mixture. Clark and Reynolds (1936) have employed the internal-standard method for estimating the amount of quartz in dusts likely to be responsible for silicosis. The method has also been used by Ballard, Oshry and Schrenk (1943).

§9. THE USE OF AN INDEX FOR IDENTIFICATION BY THE X-RAY POWDER METHOD

The extensive application of X-ray powdered-crystal photographs for identification purposes depends upon the availability of a large collection of standard patterns and ease of reference to this. Methods of identification are largely empirical, and depend upon direct comparison with standards, so that it is important that the library of photographs of standard pure substances should cover a very wide field.

One of the ways of characterizing a powder photograph is by listing the interplanar spacings, or d values, and densities corresponding with the various lines. By compiling these lists for a large number of substances, an index can be made, assuming a suitable method of arrangement is found. The most obvious way of filing is to take the spacing of the strongest line as the main determinant and place the cards in order of decreasing value of this spacing. Further ordering of the cards can be carried out by introducing sub-groups into each main group, using the spacings of the second and third strongest lines as determinants. This was the scheme suggested by Hanawalt, Rinn and Frevel (1938), and it was later adopted in the card index published by the American Society for Testing Materials.* A typical card from the A.S.T.M. file is shown in figure 14. The values of the three strongest spacings appear at the top left-hand corner, the complete list of spacings is given in the main columns, and room is provided for other useful crystallographic data when this is, or becomes, known.

* *Card File of X-ray Diffraction Data*, published jointly by A.S.T.M. and National Research Council, 1942.

To use the index for identification of an unknown substance, the spacing values of the lines on the x-ray pattern of the unknown are evaluated, and then a match is sought amongst the cards in the section of the index corresponding with the spacing of the strongest line. For positive identification it is necessary for the two lists of spacings, that of the unknown and that of the standard, to match closely over a range of values.

Besides the card index itself, an alphabetical list of the substances for which data are available is provided. This is essential to enable diffraction data on a particular substance to be found. The cards are numbered at the top right-hand corner, and these numbers are given in the alphabetical list.

As was indicated above, the value of this kind of reference catalogue increases with the amount of data it includes. The first publication included cards for thirteen hundred different substances, but, having regard to the very large number of possible chemical compounds, it will be obvious that this is not much more than a useful start on a really comprehensive file. The range of application

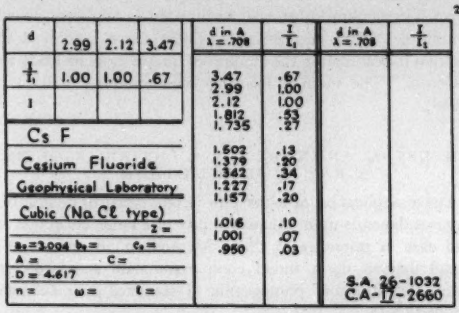


Figure 14. Card from A.S.T.M. index.

is expected to be widened by the publication through the A.S.T.M. of a supplement,* part of the data for which has been collected under the auspices of the X-ray Analysis Group of the Institute of Physics (Bragg, 1942 ; Rooksby, 1943).

However comprehensive the card index of x-ray diffraction data becomes, it is inadvisable to regard it as a substitute for careful inspection and interpretation of the photographs themselves. It is better looked upon as a valuable aid to recognition and as a means of having in one file all available diffraction data for reference purposes. Moreover, the information obtainable from powder photographs is not by any means limited to chemical identification, and the file cannot take account, for example, of line displacements resulting from solid solution effects, or of line broadening, from which much practical information on crystalline texture is obtained.

A useful survey of the limitations of the index and the difficulties that may be encountered in analysis by x-ray methods has been given by Frevel (1944). As

* *A.S.T.M. Bulletin*, March 1945, p. 37.

the index grows in size Frevel points out that the difficulties of quickly locating desired standard patterns will increase. He describes, and gives models of, a systematic classification to overcome such difficulties. The classification assumes knowledge of the elements present in the specimen whose pattern is sought, and is applicable when x-ray analysis follows preliminary spectrographic and chemical examination.

§10. THE EXAMINATION OF MIXED CRYSTALS OR SOLID SOLUTIONS

The changes accompanying the formation of mixed crystals or solid solutions that may be observed in x-ray powder photographs are of considerable practical interest. The changes are, in effect, line displacements due to the alterations in lattice dimensions involved, and measurements of such displacements can be employed to give useful analytical information.

With relatively simple solid-solution systems it is common practice to assume that, to a first approximation, a linear relation connects the crystal-lattice dimensions with atomic or molecular composition. The relation is seldom accurately true over wide ranges of composition, but is often so over narrow ranges. Even over wide ranges the approximation has been found in certain instances to be justified so long as a very high accuracy of determination is not claimed. When deviations from the straight-line relation are obviously considerably greater than the experimental limits of accuracy, the system is explored with comparatively few samples, for the points are generally found to follow a smooth curve.

The use of lattice-parameter measurements in the study of alloy phase-diagrams is well established. Phase boundaries may be placed with considerable accuracy by examining the lattice dimensional changes of a given phase that occur with composition. The lattice parameter of the phase in a two-phase region is then measured, and from the relation already established the composition at the phase boundary is evaluated. These methods have been employed extensively by the Bragg school at Cambridge (Bradley, Bragg and Sykes, 1940) and by Hume-Rothery (1944) and his collaborators at Oxford. Both binary and ternary alloy systems have been investigated by the x-ray technique.

Besides metals and alloys, solid-solution formation between compounds is common when structures are isomorphous. Oxide and other systems may be studied on similar lines to those utilized for metallic alloys, and in this field the use of x-ray methods will probably be extended. A recent example is the study of the magnesia-chrome-alumina system by W. T. Wilde and W. J. Rees (1943).

With isomorphous compounds of less simple structure, and assuming that a continuous series of solid solutions is formed, it is still practicable to base composition determinations upon planar-spacing determinations. The technique has been utilized for the examination and analysis of barium-strontium carbonate in connection with the well-known thermionic "oxide" cathode (Rooksby, 1942). In this instance x-ray methods have displaced chemical methods, providing a sufficiently accurate result with very much less expenditure of time. Chemical separation of barium and strontium is difficult, and analysis by conventional chemical procedure of a millgram or so of "oxide" coating from the cathode of a cathode-ray tube would be of doubtful accuracy.

The procedure by X rays is relatively simple. The planar spacings are evaluated for about one dozen lines on the powder photograph of the unknown carbonate. The molecular ratio of barium to strontium carbonate is then calculated by comparison of these spacings with those for pure strontium and pure barium carbonates, a value being obtained for each spacing separately. The mean value is correct to within ± 2 molecular per cent of the true value.

As with metallic alloys, the range of solid-solution formation between isomorphous compounds, when it occurs, is often discontinuous, and certainly with compounds of dissimiliar structures solid solubility is usually very limited When solid-solution changes are small, the x-ray technique must be capable of revealing very small alterations in lattice dimensions. On the x-ray powdered-crystal photograph the sensitivity of reflections to lattice dimensional changes increases with angle, and a high accuracy results when lines in the back-reflection region are utilized. This may be seen by writing the classical relation connecting planar spacing, x-ray wave-length, and angle of reflection in the form

$$\frac{n\lambda}{2} = d \sin\theta.$$

Then
$$\Delta d \sin\theta + d\cos\theta\Delta\theta = 0 \ ;$$

$$\therefore \quad \frac{\Delta\theta}{\Delta d} = -\frac{\tan\theta}{d}$$

With approach of θ towards 90°, $\tan\theta$ approaches ∞, and so any changes in d will produce larger proportionate changes in θ.

In some practical applications, all that may be required is to distinguish when solid-solution changes have occurred. Using as reference marks the fiducial shadow edges at the back-reflection end of the powder photograph, the sensitivity to small shifts in the positions of the high-order lines is high. By the use of this technique it has been shown, for example, that many impurity activators, utilized in inorganic fluorescent materials, form solid solutions of the substitutional type with the matrix compound (Jenkins, McKeag and Rooksby, 1939). Line displacements are observed as between the pure substance and the fluorescent compound, in which a small amount of the impurity activator is incorporated. Figure 15* illustrates the case of zinc orthosilicate, which gives a green luminescent colour under cathode-ray and short-wave ultra-violet excitation, when a small percentage of manganese has been added during the preparation.

§11. THE INTERPRETATION OF LINE BROADENING

(i) *Causes of line broadening*

There are two main causes of line broadening on powdered-crystal photographs which are of importance in practical applications of x-ray analysis. They may be briefly described as broadening due to small crystal size and broadening associated with variable lattice parameters. In the absence of other information the two types of broadening would not be distinguishable. It is to variations in lattice parameters within the crystal grains that distortion or strain in metals is attributable.

* Facing page 109.

Figure 13. X-ray powder photographs (Cu Kα radiation, 19 cm. camera) of various forms of hydrated and anhydrous aluminas.

(*a*) Gibbsite, $Al_2O_3.3H_2O$.
(*b*) Bayerite, $Al_2O_3.3H_2O$.
(*c*) Diaspore, $Al_2O_3.H_2O$.
(*d*) Bohmite, $Al_2O_3.H_2O$.
(*e*) Precipitated colloidal bohmite.
(*f*) γ-alumina.
(*g*) γ-alumina crystallized in presence of alkali.
(*h*) α-alumina (corundum).
(*i*) β-alumina.

To *face page* 108

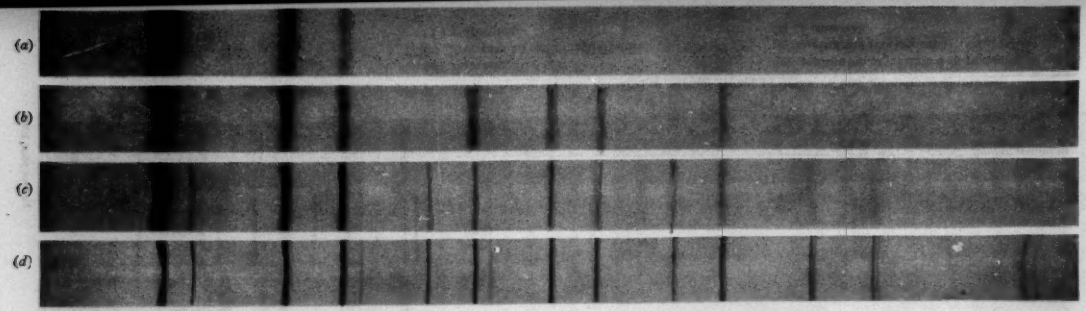


Figure 16. X-ray powder photographs (Cu Kα radiation, 19 cm. camera), comparing zinc sulphides of different ultimate particle sizes. (*a*), (*b*) and (*c*) are of precipitates prepared under various conditions ; (*d*) is of a well-crystallized reference material.

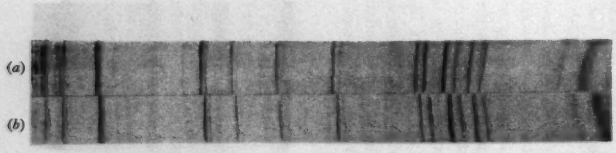


Figure 15. Parts of x-ray powder photographs (19 cm. camera, Cu Kα radiation) of
(*a*) Zinc silicate phosphor containing *c.* 2% manganese.
(*b*) Pure zinc orthosilicate.

(The prints have been sectioned down the centre to facilitate comparison of line positions.)

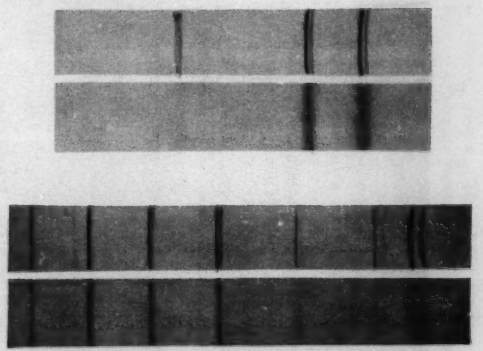


Figure 17. Parts of x-ray powder photographs indicating effects accompanying lattice distortion.

Above : Comparison between annealed powder and filings of copper.
Below : Comparison between annealed and severely ground sodium chloride.

(ii) *The estimation of crystal size from line broadening*

Broadening of the lines on a powder photograph are detectable when the average crystal size in the powder falls below about 0·5 μ. It is only detectable with certainty at this size in the high-order reflections, where the effect is noticeable as an impairment of resolution of the α doublet. As the crystal size decreases further, the line-broadening increases and the width of the lines is an inverse function of the ultimate particle size. Figure 16 illustrates the effects that are observed. The connection between line width and crystal size is generally expressed by a formula of the type

$$B = \frac{k\lambda}{t\cos\theta},$$

where B is the angular width of the diffraction line due to particle-size broadening, t is the dimension of the crystal perpendicular to the diffracting planes, θ is the angle of diffraction, and k is a constant of value approximately 0·9.

This formula was originally derived by Scherrer (1920), but has been derived independently by others (Bragg, 1933), although different values of k have been advocated. The main problem in using the X-ray technique lies in correcting line widths for instrumental broadening, so that the true value of B can be found for insertion in the Scherrer formula. The most straightforward method is the method of mixtures suggested by Jones (1938, 41), whereby a powder of large crystal size (*c.* 10^{-4} cm.) is mixed with the powder under investigation. From the lines characteristic of the standardizing substance an empirical correction curve for the geometrical broadening is plotted. The complete procedure by the method of Jones is somewhat involved, but Warren (1941), who also employs the mixture technique, suggested a simplification. All that need be done is to determine the breadth B_s of a line of the standardizing substance that occurs near a line of breadth B_m of the substance of unknown crystal size. Then the broadening β due to particle size alone is obtained from

$$\beta^2 = B_m^2 - B_s^2.$$

The breadths used by Warren are half-widths, i.e. the angular spread in radians between the points at which the intensity falls to half its peak value, but integrated breadths are preferred by Jones.

Both Jones and Warren stress that the difficulties of accurate particle-size determinations by X-ray technique are considerable, and it would be unwise to claim that the values obtained are of high accuracy. But very useful comparisons are possible, and it is relative measurements which are of chief practical importance. One virtue of X-ray determinations is that a clear distinction is obtained between true crystal size and that of aggregates or clusters of small ultimate fragments of which particles may consist.

The technique has been applied extensively by Blayden, Gibson and Riley (1943) to the study of cokes and carbons. They have examined the growth of the graphitic elements, of which "amorphous" carbons consist, during the process of coking, and have followed the changes occurring with changes in temperature of coking. They have distinguished between the rate of growth in the *c*-direction, which reflects the stacking together of the graphitic layers,

and that in the *a*-direction, which reveals the extension of the graphitic layers in their own planes.

In so far as surface area is an important factor in adsorption problems, the virtue of an extremely small ultimate crystal size possibly lies in the very large potential surface presented as compared with that of a coarse-grain structure. It is known, for example, that γ-alumina is "active" towards water vapour, whereas α-alumina is inactive. As was mentioned earlier in the discussion of figure 13, γ-alumina, as normally prepared, gives broad X-ray reflections, whereas α-alumina generally gives sharp reflections. It is thus obvious that whilst a powder consisting of α-alumina is almost invariably of a relatively coarse-grain character, a powder comprising γ-alumina can be of very small ultimate size, and is thus of practical use for application to adsorption problems.

Another example is the use of X-ray technique for distinguishing the degree of crystallization in mullite-containing refractories utilized in glass-melting tanks (Rooksby and Partridge, 1940). The X-ray powder photograph is utilized to reveal the crystal size of the mullite phase, and by means of an empirical relation the temperature of firing of the refractory can be adduced.

(iii) *Line broadening and plastic deformation.*

When a metal has been deformed beyond its elastic limit, its X-ray pattern reveals a noticeable increase in line width. There are differences in the amount of the widening as between one metal and another, but in some instances resolution of the α doublet in the high-order reflections can no longer be seen. Impairment of definition is observed for directional plastic deformation such as occurs in cold-rolling or drawing processes, and it is very pronounced for filings of some metals. Figure 17 (*a*)* shows the effects for copper. The condition responsible for such line broadening is generally described as lattice strain or distortion, although plastic strain must be distinguished from the elastic strain which occurs in the presence of stresses insufficient to cause permanent deformation.

The exact nature of the phenomena associated with lattice distortion has been the subject of much discussion. Two main theories have been presented to explain the X-ray line-broadening effect. These may be referred to as the particle-size theory and the microstress theory. One school of thought has ascribed the line broadening accompanying plastic deformation to fragmentation of the crystals to a size below the optimum that is required for sharp reflections (Wood, 1940, 41). Break-up of the crystal grains does in fact occur, but it is doubtful if this is the primary cause of line widening. The general weight of evidence from recent experimental investigations appears to be more in accord with the microstress theory. This regards the lattice strain as consisting of small interplanar distance variations from point to point in the crystal grains. These variable lattice parameters are caused by residual microstresses that remain in the lattice following the deformation processes.

Recent work has been directed to making the distinction between particle-size broadening and lattice-parameter broadening by observing the manner in which the width of X-ray reflections varies with the angle of diffraction θ. For small particle size, the variation should be a function of $\lambda \sec \theta$; this follows

* Facing page 109.

from the Scherrer relation quoted in the previous section. But if lattice-parameter variations are responsible for the increase in line width, the changes with angle should be proportional to tan θ and be independent of wave-length. Dehlinger and Kochendorfer (1939) carried out some measurements, but the results were inconclusive. More recently Smith and Stickley (1943) have described similar experiments on brass and tungsten using precision technique with monochromatized X radiation, and Stokes, Pascoe and Lipson (1943) and Lipson and Stokes (1943) have made a careful study of the effects with copper filings. The results obtained support the view that lattice-parameter variations are mainly responsible for the line broadening associated with plastic strain, although some contribution from fragmentation into small ultimate crystallites cannot be ruled out completely.

The return of line width to normal sharpness occurs when the lattice strains are eliminated. Usually the metal requires heat treatment, but with some metals recovery takes place at room temperature (Megaw, Lipson and Stokes, 1944, 45). The strain-free metal is said to be in the annealed condition, and annealing temperatures may be appreciably below those at which recrystallization or grain growth becomes visible under the microscope. X-ray measurements may be used as a criterion of adequate annealing treatment. They are so utilized, for example, for judging the best annealing conditions to be given to the nickel-iron powders employed in low-loss magnetic core applications (Rooksby, 1941). With back-reflection cameras the X-ray technique can be applied to the examination of metal specimens of a variety of sizes and forms, to reveal the presence or absence of lattice distortion. The results must, however, be carefully interpreted, since one of the limitations is that, with large metal pieces, only the surface conditons are revealed by X-ray examination, and it is not always safe to assume that these are the same as the conditions in the bulk material.

Line broadening due to lattice distortion is not a phenomenon occurring only with metals. The X-ray patterns of inorganic substances that have been subjected to severe grinding may also exhibit line-widening effects. Figure 17 (*b*)* compares back-reflection regions of X-ray powder photographs of severely ground and strain-free sodium chloride. The impairment of definition of the reflections is very marked in the case of the severely ground specimen, so that the particles have suffered considerable distortion from the grinding process. X-ray methods will no doubt become more and more valuable for this kind of examination of industrial products, because of the probable effects of plastically deformed conditions on other physical properties of the materials involved.

§12. SOME FURTHER PRACTICAL APPLICATIONS OF X-RAY ANALYSIS TECHNIQUE

(i) *The measurement of stress*

A use of X-ray analysis that appears to be growing in importance is the application to the determination of macroscopic stress in metals. If it be assumed that no permanent deformation has occurred, the presence of stresses is indicated by a departure of the inter-atomic distances from normal values. The changes in lattice spacings give a measure of the elastic strain, and from this the stress can be calculated.

* Facing page 109.

The planar spacings have to be determined with high accuracy, and for this purpose high-order reflections are employed. The use of back-reflection lines is desirable also because it is usually required to determine stresses in large metal blocks, such as castings or welded structures, and these can only be examined conveniently by back-reflection methods. Specially designed back-reflection cameras are generally utilized, the flat film cassette often being clamped to the x-ray tube and means provided for rotating it in its own plane.

For accurate interplanar spacing measurement, the specimen-to-film distance must be determined with high precision. and there must be no uncertainty from one exposure to another. Two methods of approach have been adopted. In the method employed by Thomas (1941) reliance is placed upon setting the distance very exactly, and a carefully constructed camera using locating devices and feeler gauges has been described by him. A more usual practice is to compensate for any errors in setting the distance between specimen and film by coating the surface of the specimen under examination with a standard reference substance such as gold or silver. X-ray reflections for the reference metal are then superimposed on the photograph, and from the known planar spacing values the exact specimen-to-film distance can be computed. An objection to this technique is that it weakens the reflections from the specimen itself and necessitates longer photographic exposures. Also the specimen-to-film distance is in fact very slightly shorter for the reference substance than for the material under examination, but most investigators appear to agree that the difference is so small that it can be neglected.

Since back-reflection rings of high θ value are utilized, the strain measured is that along the normal to atomic planes nearly parallel to the surface of the specimen. For example, using Co Kα radiation with an iron or steel specimen, reflection occurs from (310) planes at $\theta = 80°\ 37'$, and with aluminium and CuKα radiation the 511 reflection is obtained at approximately 81°. In the simplest case, two values of the planar spacing will be required to calculate the strain, one for the stressed condition and one for the unstressed condition. If the value of the strain is then $\Delta d/d_0$, the sum of the principal stresses in the surface of the specimen is given by

$$S = -\frac{E}{\nu}\left(\frac{\Delta d}{d_0}\right),$$

where E is Young's modulus and ν is Poisson's ratio.

It is not always practicable to determine the lattice spacing of the material in the stress-free state, and, moreover, the direction of the stress is usually required. Various methods have been employed which avoid the necessity of knowing the lattice spacing of the unstressed material. By making two exposures, one with the x-ray beam normal to the surface of the specimen and one with it inclined at, say, 45°, a value for the component of stress in a particular direction may be obtained. With the inclined exposure, the x-ray diffraction ring from the specimen is not quite circular, since different parts of the circumference of the ring originate with planes at different angles to the surface. The displacement from the reference ring of that side of the specimen ring nearest the specimen is normally measured, and the spacing of planes at a considerable angle to the surface calculated.

A single oblique exposure may also be utilized. The distances between the diffraction rings are then measured at two positions on their circumference. This again gives two values of spacings of planes differently inclined to the surface of the specimen. Methods for computing the stress component from the measurements are given by Thomas (1941) and by Barrett (1943).

The application of this x-ray technique in the light-alloy field has been described by Frommer and Lloyd (1944). In this work it was found necessary to design special cameras to deal with the difficult shapes of specimens that had to be examined. Arrangement was made for rotation or oscillation of the cassette, as, with a stationary film, the coarse-grain structure of many of the alloys gave rise to spots and streaks on the diffraction rings which hindered accurate measurement.

With other metals it is commonly desirable to move the film, although except for perpendicular exposures it is not permissible to provide for complete rotation. In the case of oblique exposures the film can only be oscillated through a few degrees of angle without impairing definition of the reflections.

X-ray equipment and cameras that are transportable will probably be essential in the future, since the x-ray method will probably have increasing application to work that cannot be moved to a fixed x-ray set. One important advantage of x-ray technique over other methods of strain measurement is that it is entirely non-destructive. The chief limitations of the method are that it requires elaborate and expensive equipment, and that it only surveys the conditions existing in the surface layers, and these may differ from those in the main body of the specimen. The results of plastic deformation must also be taken into account when elastic strain is being evaluated. Smith and Wood (1944) have shown that permanent changes in lattice dimensions can occur with plastic deformation, and consequently a single measurement of lattice parameters might give a misleading result if plastic strain were present. It is usually possible to avoid ambiguity by employing the two-exposure technique, thus avoiding any assumption about the lattice constant of the material in the stress-free condition.

(ii) *Measurements of crystal size in the microscopic range*

X-ray patterns may be used for the estimation of crystal size in the microscopic range, between 5 μ and 100 μ. A camera of the Laue type, with a flat film and the x-ray beam of about 1 mm. diameter, is a common arrangement. Both transmission and back-reflection positions for the film have been employed.

A polycrystalline specimen in a stationary position only satisfies the random requirements of the Debye-Scherrer method if the average crystal size is small enough. At about 5 μ and with increasing sizes the rings are broken and show individual reflections. Moreover, with unfiltered x rays, Laue spots begin to appear, arising from the continuous-radiation component of the x-ray beam. Further increases in crystal size emphasize the Laue spots relative to the Debye-Scherrer rings, and the size of all the reflections becomes greater.

The changes in appearance of the photographs comprising break-up of the Debye-Scherrer rings, peppering of the background with Laue spots, and changes in size of the spots, may be utilized for determining the average crystal size in the specimen examined. Clark (1933) has obtained an empirical relation between the size of the spots and crystal size for fixed conditions of x-ray examination.

In many problems it is sufficient to prepare a series of standard photographs, for comparison by visual inspection with an unknown. Examples are given in figure 18 of x-ray patterns contrasting nickel oxides of different grain sizes.

The method is particularly useful in applications in which microscopic investigation is difficult or impracticable. Jay and Chesters (1938) describe the use of the technique for the examination of the texture of magnesia refractory bricks used in the steel industry. Jay (1941 b) has also referred to x-ray evidence of a connection between the grain size of transformer sheet steel and magnetic watt-loss values. This is also reported by Clark (1940), who gives several other examples showing the value of x-ray measurements of crystal size in metallurgical problems.

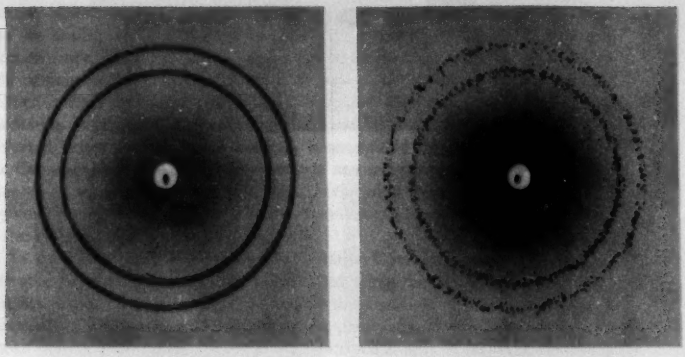
X-ray examination is utilized in many industrial applications in which texture as regards crystal size has to be correlated with heat treatment. The crystal-growth characteristics of fired refractory bodies for valve manufacture and dielectric purposes have been studied by x rays, in order to ascertain the connection between grain size and toughness, density, resistance to thermal shock, etc. (Rooksby, 1941). The inter-relation between grain size, firing temperature and impurities provides a good example of a problem whose solution is likely to be greatly assisted by the aid of x-ray analysis.

(iii) *The study of directional orientation in polycrystalline materials*

An example of preferential crystalline orientation in a polycrystalline specimen has already been mentioned. The "ordered aggregate" technique for the examination of clay minerals depends upon obtaining a specimen in which the crystal fragments are all oriented with their basal planes approximately parallel to one another. If the x-ray beam is incident in a direction nearly parallel to the plane of the specimen, the basal-plane reflections, instead of being distributed evenly around the circumference of Debye-Scherrer rings, are concentrated in special directions, and with an appropriately positioned strip of film the reflections appear considerably enhanced in intensity.

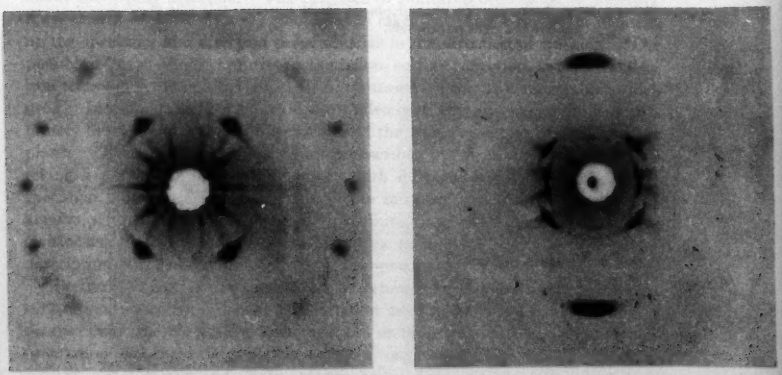
It is well known that hot and cold working of metals and alloys usually causes the crystal fragments to orient themselves in special ways with respect to the direction of working. X-ray patterns of the worked materials reveal the development of this selective orientation, the characteristic rings showing enhanced blackening or arcing at certain positions. With pronounced preferential orientations the Debye-Scherrer rings will be discontinuous around their circumference, the arcs of intense blackening being quite short. X-ray patterns of metal foils in which there is an advanced development of preferential orientation are reproduced in figure 19.

From the positions of the arcs or segments of enhanced blackening on the Debye-Scherrer rings the ideal preferred orientation of the crystallites can be specified. With drawn wires the orientation is only selective with respect to the direction of drawing; it is random about the wire axis. For advanced stages of preferential orientation in wires, therefore, x-ray diffraction patterns taken with the x rays incident perpendicular to the wire axis are similar to photo-



3 μ. 25 μ.

Figure 18. X-ray pinhole transmission patterns contrasting nickel oxides of different ultimate particle sizes.



Molybdenum. Aluminium.

Figure 19. X-ray fibre photographs (Cu radiation) showing preferred orientation in rolled metal sheets, rolling direction vertical.

To face page 114

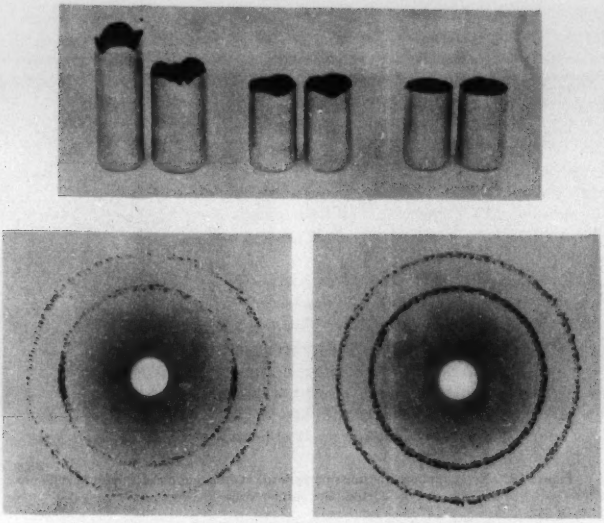
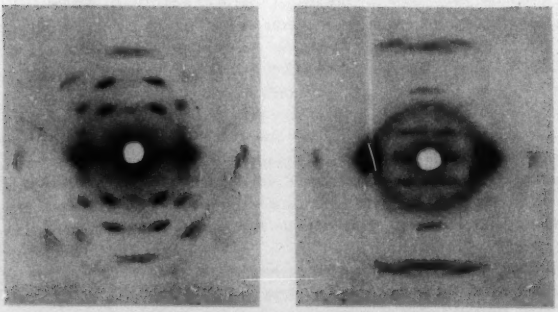


Figure 20. Preferred orientation in Cu-Ni alloy sheet.

Above: Photograph contrasting cups drawn from sheet with different degrees of preferred orientation.

Below: X-ray back-reflection photographs of sheets corresponding with cups on the left and right respectively.



(*a*) (*b*)

Figure 21. X-ray fibre photographs (Cu Kα radiation, 4·0 cm.) of (*a*) a well-oriented regenerated cellulose, (*b*) nylon.

graphs of a single crystal rotated about a principal axis. If the X-ray beam is incident along the wire axis, the Debye-Scherrer rings show no arcing.

With wires, the orientation developed is often referred to as a *fibre structure* because of its resemblance to the structure of organic fibrous materials. Few data are available which suggest that selective orientation in wires is of major practical importance. There is evidence that molybdenum and tungsten wires with pronounced preferred-orientation texture have increased tensile strength and reduced elongation on annealing as compared with wires in which selective orientation is feebly developed. Burghoff and Porter (1944) have found that preferred orientation in brass wires is accompanied by both a reduction in ductility and increases in yield and tensile strengths over values in randomly oriented wire of similar grain size. Resistivity may also be influenced by the degree of preferred orientation, so that X-ray measurements may be a useful guide to the assessment of these various qualities.

In rolled materials, the preferred-orientation textures appear to be of considerable practical significance. With rolled strip the crystals tend to align themselves not only with a specific axis in the direction of rolling, but also in a particular orientation with respect to the plane of rolling. For instance, in drawn molybdenum wire, the texture is described simply in terms of the alignment of the [110] direction, the direction parallel to a cube-face diagonal, with the wire axis. In rolled molybdenum sheet there is an additional restriction on the orientation, and planes parallel to a set of cube faces tend to become parallel to the plane of the sheet. The ideal preferred orientation is thus described as a (100) [110] texture.

Usually there is a substantial spread in orientation of the crystallites about the ideal preferred arrangement. In sheet or strip this is normally most marked about the rolling direction. The presence of preferred orientation is therefore most clearly revealed on X-ray photographs taken with the X-ray beam incident on the specimen in a direction perpendicular to the direction of working. The possible X-ray methods comprise transmission, surface reflection, and back-reflection techniques, though for rolled strip, transmission methods have mostly been used. In order to define the texture completely in any given example it is customary to use pole figures of one or other of the more important crystallographic planes. The mapping of a pole figure as described by Barrett (1937, 43) involves taking a number of X-ray photographs with different angular settings of the specimen. There will be variations in the intensity maxima for the different angular settings, depending upon the distribution of orientations, and these are transposed to the pole figure, which in its final form provides a complete description of the orientation texture in the specimen.

Preferred-orientation textures may be retained on heat treatment or recrystallization, or an entirely fresh orientation texture may result. With body-centred cubic metals, such as molybdenum or tungsten, the original preferred-orientation condition is generally retained on recrystallization, but in face-centred cubic metals and alloys, copper and α-brass for example (Cook and Richards, 1940), a new orientation is commonly formed on annealing. On the other hand, if the directional orientation is only feebly developed it may disappear completely on recrystallization.

One or two practical implications of preferred orientation in rolled materials may be mentioned. It is found that the formation of "ears" on cups made by deep-drawing operations is attributable to the preferred orientation of the crystals existing in the strip or sheet from which the cups are fabricated (Cook, 1937). An example of the effect is shown in figure 20. The cups were fabricated from two different consignments of rolled sheet of a copper-nickel alloy. The x-ray photographs shown in figure 20 reveal that the batches of metal from which the eared cups were made have a strongly marked preferred-orientation texture which is absent from the other material. Rolling schedules can usually be adjusted to prevent the development of a too strongly marked preferred orientation, and x-ray technique can provide a means of control of the processes involved.

The effect of cross-rolling (rolling in two directions at right angles) also has important practical results. In a body-centred cubic metal such as molybdenum the (100) [011] texture is prominent, but with uni-directional rolling there is usually a considerable scatter about the rolling direction. If, however, after some proportion of the reduction has been effected, the rolling direction is altered by 90°, this scatter about the first rolling direction tends to disappear. The ideal (100) [011] texture is thus closely approached, and this means that (100) cleavage planes are perpendicular to the plane of the rolled sheet and at 45° to the rolling directions. Such sheet exhibits marked brittleness in the 45° directions (Ransley and Rooksby, 1938).

Strongly marked directional orientation may be required in some applications to confer some specific property on the rolled material. For example, by carefully controlled rolling processes, Goss (1935) has shown that it is practicable to produce a strongly marked preferred-orientation texture in silicon-iron sheet for transformer cores. This texture is such that the direction of easiest magnetization of the silicon-iron crystals, parallel to the cubic axis, coincides with the rolling direction. The sheet consequently has a high permeability in the rolling direction (Bozorth, 1935).

Directional orientation of the crystals is also found in electro-deposited metals, and in volatilized and sputtered films of both metallic and inorganic chemical materials. Wood (1935, 37) has studied electro-deposited nickel and chromium by x-ray methods, and has found that the brightness of the deposit depends upon the degree of directional orientation present. Similar results have also been described by Hume-Rothery and Wyllie (1943) for electro-deposited chromium.

Enough has been said to show that directional or preferred orientations in industrial materials are of considerable practical importance. X-ray examination provides a possible means both for controlling the products at different stages of the manufacturing process, so as to ensure a consistent material, and for ascertaining that the finished material has either the isotropic or anisotropic properties that are required for the particular application.

It remains to be added that x rays have been applied extensively to the study of natural and synthetic organic fibres in which molecular orientation is prominent. In the natural fibres, bundles of long-chain molecules extend along the axis of the fibres, and in many regions the regularity of alignment permits the development of a crystalline arrangement. The molecules themselves may

extend through several crystalline regions, intervening parts of the fibre being non-crystalline because, in these parts, the molecules are not uniformly oriented. The crystallites have, in general, similar orientations with respect to the fibre axis, and these give rise to the well-known fibre photograph. By analogy with single-crystal rotation photographs much structural information has been deduced from detailed analyses of fibre photographs. Many of the fundamental principles and concepts in this subject are due to the work of Astbury and his associates (Astbury, 1945).

Some of the more obvious characteristics of the photographs have been utilized in a practical way for assessing the effects of different treatments of the fibres. There is evidence that the strength of the fibre can be correlated with the perfection of orientation of the crystallites, so that methods of treatment, for example drawing and extrusion, which would be expected to emphasize and develop the preferential orientation, are utilized to obtain high tensile strengths. Figure 21 (*a*) shows the X-ray pattern of an artificial cellulose fibrous material, and figure 21 (*b*) that of a typical nylon thread. In both instances threads of considerable strength have been produced by extrusion or drawing processes that cause the crystallites to orient themselves very uniformly along the axis of the thread. Extended drawing of the fibres has the additional advantage that in improving molecular alignment it inevitably decreases the proportion of "amorphous" regions, where in all probability the greatest source of weakness lies.

REFERENCES

VAN ARKEL, A. E., 1926. *Physica*, **6**, 64.

ASTBURY, W. T., 1940. *Textile Fibres under the X-rays* (London ; Imperial Chemical Industries).

ASTBURY, W. T., 1945. *Chemistry and Industry*, 114.

ASTBURY, W. T. and PRESTON, R. D., 1934. *Nature, Lond.*, **133**, 460.

BALLARD, J. W., OSHRY, H. I. and SCHRENK, H. H., 1943. *J. Opt. Soc. Amer.*, **33**, 667.

BARRETT, C. S., 1937. *Metals Technology Publication*, No. 819.

BARRETT, C. S., 1943. *The Structure of Metals*, chap. 14 (New York : McGraw-Hill).

BISCOE, N. and WARREN, B. E., 1942. *J. Appl. Phys.*, **13**, 364.

BLAYDEN, H. E., GIBSON, J. and RILEY, H. L., 1943. *Conference on Ultra-fine Structure of Coals and Cokes* (London : B.C.U.R.A.).

BOUWERS, A., 1930. *Physica*, **10**, 125.

BOZORTH, R. M., 1935. *Trans. Amer. Soc. Metals*, **23**, 1107.

BOZORTH, R. M. and HAWORTH, F. E., 1938. *Phys. Rev.*, **53**, 538.

BRACKNEY, H. and ATTLEE, Z. J., 1943. *Rev. Sci. Instrum.*, **14**, 59.

BRADLEY, A. J., BRAGG, W. L. and SYKES, C., 1940. *J. Iron Steel Inst.*, **141**, 63.

BRADLEY, A. J. and JAY, A. H., 1932. *Proc. Phys. Soc.*, **44**, 563.

BRADLEY, A. J., LIPSON, H. and PETCH, N. J., 1941. *J. Sci. Instrum.*, **18**, 216.

BRAGG, Sir LAWRENCE, 1942. *Nature, Lond.*, **150**, 738.

BRAGG, W. H. and BRAGG, W. L., 1933. *The Crystalline State* (London : G. Bell and Sons).

BROWN, K. W., CHIRNSIDE, R. C., DAUNCEY, L. A. and ROOKSBY, H. P., 1944. *G.E.C. Journal*, **13**, 53.

BUERGER, M. J., 1942. *X-ray Crystallograpny* (New York : John Wiley and Sons).

BUNN, C. W., 1941. *J. Sci. Instrum.*, **18**, 70.

BURGHOFF, H. L. and PORTER, J. S., 1944. *Metals Technology Publication*, No. 1688.

CLARK, G. L., 1933. *Chemistry and Industry*, **52**, 336.

CLARK, G. L., 1940. *Applied X-rays* (New York : McGraw-Hill).

CLARK, G. L. and REYNOLDS, D. H., 1936. *Ind. Engng. Chem.* (*Anal. Ed.*), **8**, 36.
CLAUSSEN, G. E. and SKEHAN, J. W., 1942. *Metals and Alloys*, **15**, 599.
CLAY, R. E., 1934. *Proc. Phys. Soc.*, **46**, 703.
COOK, M., 1937. *J. Inst. Met.*, **60**, 159.
COOK, M. and RICHARDS, T. LL., 1940. *J. Inst. Met.*, **66**, 1.
DARWIN, C. G., 1914. *Phil. Mag.*, **27**, 315 and 675.
DARWIN, C. G., 1922. *Phil. Mag.*, **43**, 800.
DAVEY, W. P., SMITH, F. R. and HARDING, S. W., 1944. *Rev. Sci. Instrum.*, **15**, 37.
DEBYE, P. and SCHERRER, P., 1916. *Phys. Z.*, **17**, 277.
DEBYE, P. and SCHERRER, P., 1917. *Phys. Z.*, **18**, 291.
DEHLINGER, U. and KOCHENDORFER, A., 1939. *Z. Metallk.*, **31**, 231.
DONNAY, J. D. H. and FANKUCHEN, I., 1944. *Rev. Sci. Instrum.*, **15**, 128.
DUMOND, J. W. M., WATSON, B. B. and HICKS, B., 1935. *Rev. Sci. Instrum.*, **6**, 183.
EDWARDS, OLIVE S and LIPSON, H., 1941. *J. Sci Instrum.*, **18**, 131.
FRANKEL, J. J., 1944. *J. Chem. Soc. S. Afr.*, **44**, 169.
FREVEL, L. K., 1944. *Industr. Engng. Chem.* (*Anal. Ed.*), **16**, 209.
FRICKE, R. and others, 1941. *Z. Elektrochem.*, **47**, 374.
FROMMER, L. and LLOYD, E. H., 1944. *J. Inst. Met.*, **70**, 91.
GOSS, N. P., 1935. *Trans. Amer. Soc. Met.*, **23**, 511.
DE GRAAF, J. E. and OOSTERKAMP, W. J., 1938. *J. Sci. Instrum.*, **15**, 293.
GRENINGER, A. B., 1935 a. *Trans. A.I.M.E.*, **117**, 61.
GRENINGER, A. B., 1935 b. *Z. Krist.*, **91**, 424.
GUINIER, A., 1939. *Ann. Phys., Paris*, **12**, 161.
GUINIER, A., 1943. *J. Chim. phys.*, **40**, 133.
HANAWALT, J. D., RINN, H. W. and FREVEL, L. K., 1938. *Industr. Engng. Chem.* (*Anal. Ed.*), **10**, 457.
HENDRICKS, S. B. and ALEXANDER, L. T., 1939. *Soil Sci.*, **48**, 257.
HULL, A. W., 1917 a. *Phys. Rev.*, **9**, 84.
HULL, A. W., 1917 b. *Phys. Rev.*, **9**, 564.
HULL, A. W., 1917 c. *Phys. Rev.*, **10**, 661.
HUME-ROTHERY, W., 1944. *The Structure of Metals and Alloys* (*Institute of Metals Monograph*, No. 1).
HUME-ROTHERY, W. and WYLLIE, M. R. J., 1943. *Proc. Roy. Soc.*, A, **181**, 331.
HUTTON, C. O. and FANKUCHEN, I., 1938–40. *Mineral Mag.*, **25**, 201.
IMHOFF, C. E. and BURKARDT, L. A., 1943. *Amer. Soc. Test. Mat.*, Reprint No. 93.
JAY, A. H., 1941 a. *J. Sci. Instrum.*, **18**, 128.
JAY, A. H., 1941 b. *J. Sci. Instrum.*, **18**, 83.
JAY, A. H. and CHESTERS, J. H., 1938. *Trans. Ceram. Soc.*, **38**, 209.
JENKINS, H. G., MCKEAG, A. H. and ROOKSBY, H. P., 1939. *Nature, Lond.*, **143**, 978.
JONES, F. W., 1938. *Proc. Roy. Soc.*, A, **166**, 16.
JONES, F. W., 1941. *J. Sci. Instrum.*, **18**, 157.
KIRKPATRICK, P., 1941. *Phys. Rev.*, **59**, 452.
KRATKY, O., 1942. *Naturwissenschaften*, **30**, 542.
LAUBENGAYER, A. W. and WEISZ, R. S., 1943. *J. Amer. Chem. Soc.*, **65**, 247.
LIPSON, H. and STOKES, A. R., 1943. *Nature, Lond.*, **152**, 20.
LIPSON, H. and WILSON, A. J. C., 1941. *J. Sci. Instrum.*, **18**, 144.
LONSDALE, K., 1942. *Proc. Phys. Soc.*, **54**, 314.
LONSDALE, K., 1942–43. *Rep. Prog. Phys.*, **9**, 256.
LUKESH, J. S., 1940. *Rev. Sci. Instrum.*, **11**, 200.
MACARTHUR, I., 1944–45. *Electronic Engng.*, **17**, 272 and 317.
MAZELEO, L. YA., 1940. *J. Appl. Chem. U.S.S.R.*, **13**, 1288.
MAZELEO, L. YA., 1943. *Ceramic Abstr.*, **22**, 42.
MEGAW, H., LIPSON, H. and STOKES, A. R., 1944. *Nature, Lond.*, **154**, 145.
MEGAW, H. and STOKES, A. R., 1945. *J. Inst. Met.*, **71**, 279.
MÜLLER, A., 1931. *Proc. Roy. Soc.*, A, **132**, 646.
MÜLLER, A. and CLAY, R. E., 1939. *J. Inst. Electr. Engrs.*, **84**, 261.
NAGELSCHMIDT, G., 1941. *J. Sci. Instrum.*, **18**, 100.
NAGELSCHMIDT, G. and KING, E. J., 1941. *Bio-chem. J.*, **35**, 152.
NELSON, J. B. and RILEY, D. P., 1945. *Proc. Phys. Soc.*, **57**, 160.

OWEN, E. A. and PICKUP, L., 1932. *Proc. Roy. Soc.*, A, **137**, 397.
RANSLEY, C. E. and ROOKSBY, H. P., 1938. *J. Inst. Met.*, **62**, 205.
RILEY, D. P., 1943. *Conference on Ultra-fine Structure of Coals and Cokes* (London: B.C.U.R.A.).
RIVLIN, R. S., 1940. *Nature, Lond.*, **146**, 806.
ROBERTSON, J. M., 1943. *J. Sci. Instrum.*, **20**, 175.
ROBERTSON, R., FOX, J. J. and MARTIN, A. E., 1934. *Phil. Trans.*, A, **232**, 463.
ROBERTSON, R., FOX, J. J. and MARTIN, A. E., 1936. *Proc. Roy. Soc.*, A, **157**, 579.
RONNEBECK, H. R., 1943. *J. Sci. Instrum.*, **20**, 154.
ROOKSBY, H. P., 1929. *Trans. Ceram. Soc.*, **28**, 399.
ROOKSBY, H. P., 1939. *J. Soc. Glass Techn.*, **23**, 76.
ROOKSBY, H. P., 1941. *J. Sci. Instrum.*, **18**, 84.
ROOKSBY, H. P., 1942. *Electrical Times*, **102**, 185.
ROOKSBY, H. P., 1943. *Electrical Times*, **104**, 212.
ROOKSBY, H. P., 1945. *J. Soc. Glass Techn.*, **29**, 258.
ROOKSBY, H. P. and PARTRIDGE, J. H., 1940. *J. Soc. Glass Techn.*, **24**, 110.
RYDE, J. W. and YATES, D. E., 1926. *J. Soc. Glass Techn.*, **10**, 274.
SCHERRER, P., 1920. See R. ZSIGMONDY'S *Kolloidchemie*, 3rd Edition (Leipzig: Spamer), 387.
SMITH, C. S. and STICKLEY, E. E., 1943. *Phys. Rev.*, **64**, 191.
SMITH, S. L. and WOOD, W. A., 1944. *Proc. Roy. Soc.*, A, **182**, 404.
STOKES, A. R., PASCOE, K. J. and LIPSON, H., 1943. *Nature, Lond.*, **151**, 137.
TATEL, H. and HULTGREN, R., 1938. *Rev. Sci. Instrum.*, **9**, 47.
TAYLOR, A., 1941. *J. Sci. Instrum.*, **18**, 91.
TAYLOR, A. and SINCLAIR, H., 1945. *Proc. Phys. Soc.*, **57**, 126.
THOMAS, D. E., 1939. *J. Sci. Instrum.*, **16**, 222.
THOMAS, D. E., 1941. *J. Sci. Instrum.*, **18**, 135.
WARREN, B. E., 1936. *Phys. Rev.*, **49**, 885.
WARREN, B. E., 1941. *J. Appl. Phys.*, **12**, 375.
WEISER, H. B. and MILLIGAN, W. O., 1934. *J. Phys. Chem.*, **38**, 1175.
WILDE, W. T. and REES, W. J., 1943. *Trans. Ceram. Soc.*, **42**, 123.
WOOD, W. A., 1935. *Phil. Mag.*, **20**, 964.
WOOD, W. A., 1937. *Phil. Mag.*, **24**, 772.
WOOD, W. A., 1940. *Proc. Phys. Soc.*, **52**, 110.
WOOD, W. A., 1941. *J. Sci. Instrum.*, **18**, 153.
WOOSTER, N., 1943. *Industrial Diamond Review*, **3**, 1.
WOOSTER, W. A. and MARTIN, A. J. P., 1936. *Proc. Roy. Soc.*, A, **155**, 150.

SOUND

By E. G. RICHARDSON
Kings College, Newcastle-on-Tyne

§1. THE PRESENT POSITION IN AUDITORIUM ACOUSTICS

In the early years of the century, and, indeed, fitfully through the preceding century, attempts to improve the acoustics of auditoria were directed in the main to getting an acceptable time of reverberation for tones in the middle of the musical scale. It is well known that W. C. Sabine of Harvard first attacked this question in a truly systematic way, altering one factor at a time, and that he finally evolved a technique—and a formula—by means of which an architect could design a hall to have an acceptable reverberation. The Sabine formula has stood the test of thirty years of application with but slight modification, but it has become apparent of late that getting the correct time of reverberation for a frequency about 500 c./sec. is not all that is necessary to give a building "good acoustics". Sabine himself recognized this when he adopted the schlieren and the ripple-tank methods to observe where the sound waves in model sections of buildings went after being sent out; he also repeated his reverberation measurements over five octaves.

As regards the latter point, it is now the custom to quote absorption coefficients for materials and to calculate reverberation times at three frequencies round about 100, 500 and 4000 c./sec. to represent the extreme bass and treble as well as the middle frequencies, but research is still in progress to decide what are the desirable reverberation and frequency characteristics for auditoria designed for different purposes. It is recognized that the curves for speech and music should not be the same. Whereas the musician likes strong reinforcement in the bass, the intelligibility of speech often depends on the amount of high frequency that the listener picks up, since high frequencies are prominent in the transients which characterize most of the consonant sounds. In broadcast studios, however, it is generally thought that the ideal to be aimed at, at least when speech and chamber music are being radiated, is an acoustic condition to give the listener the impression that the music or speech is originating in his own room. This requirement demands rapid damping of sounds in the studio, except that the treble may, with advantage to clarity, be more lingering when the voice is being broadcast. In small studios these conditions would obtain without the necessity for much correction, but something has to be done whenever a few players or speakers are grouped in a large studio.

In Sabine's theory it is assumed that sound decays exponentially in a room. Research by Mason and Moir (1941) has shown that, in those having large protuberances such as galleries comparable with, or larger in average dimension than, the wave-length of the sound, the decay is markedly discontinuous, often having a series of peaks of intensity. These workers were able to trace out the sound envelope as received by a microphone from the loud-speakers in a cinema to which a short wave-train of constant frequency has been supplied. Three

traces from their records are shown in figure 1 : (*a*) in a theatre well-lined with absorbent so that all reflections are weak and only the primary impulse stands out ; (*b*) on one of moderately good acoustics having some fairly large reflecting surfaces ; (*c*) in a simple box-shaped room having hard parallel surfaces producing strong reflections, some of which interfere favourably at the microphone to produce peaks higher than the direct radiation.

They also point out that the sound heard by the ear depends greatly on the form of the reverberation-time curve in the first few milliseconds of the decay, that is, up to the instant when it begins to be masked by the oncoming next syllable of speech or note of music. It is important that any reflected sounds falling on the ear at this period should either come so hot on the heels of the direct sound as to add to it without loss of intelligibility or be quite absent if open-air conditions are looked for.

This brings us to the question of prominent reflections. If these are established between hard and parallel surfaces, standing waves will be set up whenever a natural frequency of the air column is struck. Even "off the note", disturbance of the average reverberation may be expected, so that such a room will have a much indented reverberation-frequency curve. Moreover, what are known as

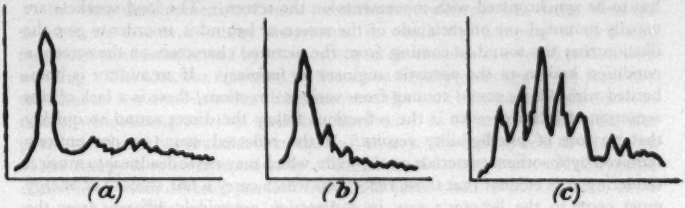


Figure 1.

"flutter echoes" may follow the emission of a pulse of sound between the two walls, as the ear is bombarded from left and right alternately with periodic waves having slowly decaying intensity. In a small room, these echoes can turn a pulse into a musical note, just as a fence of overlapping slats can turn the sound of a nearby footfall into a note ; in a larger room the impulse will be overlaid with a waxing and waning of intensity as its sound dies away (Maa, 1941).

Two remedies for this state of affairs are available. One is to cover the offending walls with absorbent, but not to an extent unjustified by the requisite reverberation time. The other involves inclining the walls at a small angle of about 10 degrees or attaching panels similarly skewed to them. Rather unsightly in an auditorium, the latter expedient is commonly adopted in small studios for recording and broadcasting (Content and Green, 1944).

In any event, the best treatment for most auditoria is to diffuse the sound outside of the stage, regarding this as a radiator and, therefore, providing it with hard surfaces to send out the energy into the body of the hall. The absorbent, if any, required to give a reasonable amount of reverberation is then distributed fairly evenly over the portion occupied by the audience. The convex reflecting surface is a device to aid general diffusion. Such surfaces are often set vertically

at the sides of the dais or proscenium arch—cf. the former Queen's Hall, London—but there is now a suggestion to set half-cylinders with axes horizontal on the ceilings and along the side walls of concert halls. Such polycylindrical reflectors have already appeared in studios for recording and broadcasting. Volckmann (1942) and Boner (1942) describe experimental studios in the United States where reflectors of quarter-inch plywood bent into arcs of about 3 ft. radius have been successfully used. The latter author prefers to leave spaces for flat surfaces between the cylinders, or to use different radii of curvature in adjacent cylinders, as, if the pattern is repeated all over the room, the structure acts selectively on wave-lengths of sound approximating to the width of the reflector; it acts in fact as a diffraction grating. As a matter of experience, Boner finds that the decay in rooms so contructed approximates more closely to a steady exponential fall without peaks than in the conventional room with flat or steadily curved walls. In this construction, rooms up to 8000 cu. ft. in volume may have a uniform reverberation time for frequencies from 40 to 17,000 c./sec. Furthermore, with a well-diffused sound pattern, the search for the best microphone position, which wasted a lot of the sound engineer's time, is obviated.

Other problems arise in large cinemas, where the sound from the loud-speaker has to be synchronized with movements on the screen. The loud-speakers are usually mounted one on each side of the screen or behind it, in order to give the illusion that the sound is coming from the pictured characters on the screen, a condition known to the acoustic engineer as *intimacy*. If an auditor is bombarded with diffuse sound coming from various directions, there is a lack of this sensation of intimacy even if the reflections follow the direct sound so quickly that no loss of intelligibility results. If the reflected sound is not entirely removed by absorbent materials on the walls, which may cause deadness to musical rendering, it is evident that those reflections which carry a fair amount of energy must come to the listener's ears in a direction not widely different from the direct sound. The condition for intimacy is then that the prominent reflected ray must subtend with the direct one as small an angle as possible. This is to be secured by having a low ceiling and by damping any rays that might come off the side walls near the screen or off the rear wall by the application of absorbent material to these parts (Mason and Moir, 1941).

As there are now suggestions to locate the source of sound at a definite point on the screen, and thence to move it about following the image for stereophonic effects (Fletcher, 1941), quite complex problems in applied acoustics will have to be considered when to a phase difference between the sounds from two loud-speakers are added the phase differences due to reflections.

In a paper by Hill (1942), the use and abuse of the public-address system in large halls is discussed. Two systems are in use. In the *concentrated* system, a cluster of loud-speakers is mounted at a considerable height above the speaker in view of all members of the audience, so that each receives direct sound. The problem is then the same as that in a hall with a fixed source of high intensity. In the *distributed system* there is one loud-speaker for each group of the audience. This is mounted a few feet above their heads, pointing slightly downward. In the latter case, the intention is that each group shall have its own local source of sound, of which it absorbs—literally—the whole output, and is out of range of

all the other speakers, including the prime (human) mover. The loud-speakers are, if possible, mounted on pillars or similar obstructions on the side remote from the platform or pulpit, so that the pillar casts a sound shadow to prevent sound seeping back towards the source.

The distributed system is better in a very reverberant building such as a cathedral, where mere amplification of a speech would result in cacophony. It is far better to relay the sound in such directions and at such a level of intensity that the subdued reverberation of the unabsorbed tones around the vault has little effect on the intelligibility at floor level. The speaker must then hold his voice at conversational level. The only disadvantage is that for the majority the sound does not appear to come from the speaker himself. Loud-speakers feeding groups of seats unoccupied on any occasion are turned off.

A technical point arises in the use of the public-address system, in the prevention of feed-back to the microphone and from one loud-speaker to another. In spite of the cunning use of acoustic shadow, this is difficult to secure with tones in the bass, since diffraction is easier for these. It is usually necessary to cut off or subdue the lower frequencies by filters in the amplifier circuits, but their absence is not detrimental where speech only is concerned.

In the development of new absorbents, research has been directed to what are known as cavity absorbents, in which the damping of air-borne transient sound is facilitated by resonance in an air vessel or in a vessel stuffed with wadding. The idea was first tried by Stoney some fifty years ago in the form of a textile sheet mounted at a short distance from the wall. In the modern developments, wire netting of close mesh has been tried in front of a flat wall, also pyramids of rock wool attached to the wall so as to leave air spaces in which the sound energy is to be frittered away. Others have formed a sort of honeycomb of textile or sheet asbestos covered with a flat sheet which is perforated to allow the sound to enter the cavities (Sabine, 1942).

This type of absorbent depends greatly for its success on the sound entering the throats of the cavities at perpendicular incidence. Rayleigh showed that a porous body consisting of capillary cavities rapidly loses its absorbing power as grazing incidence is approached. In consequence they are best mounted on a wall facing the source of sound so as to trap the first wave arriving normally. If mounted, for instance, on the ceiling of a low-built room, so that the direct rays arrive at grazing incidence, they lose their efficacy (Beranek, 1942 ; Zwikker and Kosten, 1941).

This type of material shows considerable promise as a "black-body" absorbent for sound, of which more may be heard in the future.

§ 2. METALLURGICAL APPLICATIONS OF ULTRASONICS

The science of ultrasonics is finding increasing application in the examination and preparation of ingots. The most noteworthy of these is the detection of hidden flaws and cracks. The first to employ ultrasonics for this purpose was Sokolov (1935 a). He used a source attached to one end of the specimen and passed the radiation on its emergence through a liquid (xylol) in which the moving compressions acted like the bars in a mobile diffraction grating and could be

examined by passing a beam of monochromatic light athwart the trough of liquid (Brillouin effect). According to the clarity of the optical diffraction spectrum produced by the waves in different parts of the field, he judged whether the specimen were truly homogeneous or subject to faults which cut off the ultrasonic radiation. A similar method has been used more recently by Pohlman (1939), but for detecting the radiation he uses a suspension of aluminium in xylol. The flakes set themselves normal to the ultrasonic wave-front like a set of miniature Rayleigh discs. Places in the acoustic shadow of holes or cracks in the metallic specimen under examination are detected by noting with the aid of a light beam where the orientation of the flakes is still incoherent.

A more direct method can be used. If an ultrasonic emitter is clamped to one end of a bar and the receiver to the other, the intensity of the transmitted radiation can be measured by a valve voltmeter in a circuit coupled to the latter. By comparing whole and imperfect samples of concrete beams, Meyer and Buchmann (1938), using a magneto-strictive source of 22·5 kc./sec., were able to show that faults could be detected by such an apparatus. (Behr, 1943 a, says that Mühlhauser was the first to do this—reference not quoted.) The difficulty in such comparison methods is in securing equality in the coupling between the radiator and sender on the one hand and substituted specimens on the other. A film of water or of oil is often used to ensure continuity without air gaps. A system has been developed by the firm of Hughes in this country in which a short train of waves is sent out and its wave-form as recorded by the receiver noted on a cathode-ray oscillograph. They use quartz oscillators, sometimes one at each end of an ingot (transmission method), sometimes on two facets of a prismatic specimen (reflection or echo-sounding method). The presence of hair-line cracks is shown in the latter method by a number of secondary echoes which reach the second quartz before the main reflection from the base of the prism. Whereas large fissures may be more readily detected by x rays, the ultrasonic method is best for spotting these tiny cracks, since it needs but the narrowest break in the continuity of the metal to give rise to a secondary reflection.

Ultrasonics may profitably be used, too, in the examination of specimens which are free from faults, to give information about their elastic properties. In one such application, due to Ludloff (1940), a disc of material is cut and stuck to an ultrasonic emitter so that the two form a compound plate. Chladni figures may be formed on the plate when the combination is set in oscillation, and from their configuration—after some intensive calculation!—the elastic moduli of the upper disc may be found. As an alternative to dust figures, the pouring of a film of oil on the surface and a photograph of the crispations formed in the film by the vibration have been suggested. Siegel (1944) attaches the specimen, in the form of a rod, to a similar rod of quartz, and sets the combination in oscillation by exciting the latter piezo-electrically. The natural frequencies of the composite rod are determined from the reaction on the driving circuit as the tuning of the latter is continuously altered. He was able thus to investigate the variation of the elastic and damping properties of single and polycrystalline samples of nickel with the intensity of magnetization and temperature and the change in elastic moduli of copper-gold alloys with temperature. Information from such experiments advances our knowledge of the solid state.

Bär (1940) uses a different method, in some respects like the first Sokolov research. He cuts away one end of a solid rod to a bevel of 45° and attaches a quartz oscillator to the bevelled surface. When this system is immersed in a liquid, waves which actually result from a combination of transverse and longitudinal ultrasonic waves within the solid material emerge from the sides of the rod. These waves in the liquid are examined by an optical *schlieren* method. A theory is given to show how the ultrasonic field is related to the elastic constants of the specimen.

Other applications of ultrasonics in metallurgy concern the early discovery that the high-frequency energy can influence what happens during the cooling and solidification of molten alloys. The first of these applications was to de-gas the melt, namely, to remove bubbles of gas formed by chemical reaction which would otherwise remain occluded in the solid and spoil its continuity. Sokolov (1935 b) found that the radiation also had a grain-refining effect on a melt of zinc, and Schmidt and his collaborators (1937, 39) have found that when antimony solidifies it does not exhibit its usual tendency to supercool if kept so irradiated. In duralumin similarly treated, the grain boundaries, which indicate separation of the alloy constituents, did not form, while in silumin, needle-like crystals of silicon which tend to separate out are broken up into smaller particles. Such effects are probably due to the setting up of frictional forces between the molten metal and the crystallites which come out of it during solidification.

Behr (1943 b) suggests that ultrasonics might be employed for grain-refining purposes during spot welding, the radiation being led into the work through one of the electrode faces or through a third electrode.

Other uses in metallurgy which are either being tried or considered are (i) the prevention of the formation of an oxide coating during the soldering of aluminium by agitation of the molten solder; (ii) the acceleration of the nitriding process in the hardening of steel; (iii) the hastening of solid diffusion processes—it is already known that ultrasonic agitation much accelerates diffusion in fluids; (iv) the stabilization of hardness fluctuations in tool steels (Behr, 1943 a).

The principal deterrent to the general adoption of these methods in industry—apart from those dealing with fault-finding and the testing of specimens for elasticity—is that most of them have not passed beyond the laboratory stage. It is no easy matter to bathe a large volume of solid or liquid in ultrasonic radiation, since the operating surface of the oscillator is usually small and it is in the nature of the high-frequency radiation that it travels as a beam confined nearly to a cylindrical volume erected on this small surface as base. Large emitters such as the composite ones used at sea in hydrographical work will need to be applied to these processes before they can be employed in industrial operations turning out a large mass of ingots every day.

§3. MUSICAL SCALES AND TUNING

Quite a lot of thought seems to have been given lately both by musicians and physicists to problems of temperament and tuning. Perhaps, like the theory of dimensions, it has been a soothing subject for home contemplation during the blacked-out evenings!

It is well known that the necessity which musicians feel to modulate into different keys during the course of a composition makes demands on a keyboard instrument with at least one sound source per note which it is practically impossible to fulfil if in each key the correct pitch relationships of our scales is to be preserved. The compromise through which certain notes of the scale are put out of tune in order that a number of keys may be played in without violent deviations from the correct intervals in any one of them is known as *temperament*. The twelve-note scale in the octave of the music of Western nations involves two kinds of tone, major (9/8) and minor (10/9) as well as the semi-tone (16/15).

In general the relative positions of these in a scale must be preserved. The second-degree note, however, must be liable to adjustment if it is always to sound in tune—for example, if it is to make a true fifth with the sixth note of a major scale. It becomes a mutable note ; such are characteristic of the intonation of modal music (Lloyd, 1943). The compromise tuning known as *mean-tone temperament* was originally devised for modal music, and it substituted a compromise whole tone for both major and minor tones, preserved the true relationship of the notes in a major third, but reduced the perfect fifths by building them up from semi-tones and mean tones (1.118), to which both kinds of tone were approximated (White, 1943). It has been pointed out by a number of arithmetically inclined musicians that a keyboard with 53 notes to the octave could give as close an approximation to the just scale as one could expect the ear to accept without question (Dufton, 1941, 42), but mean-tone temperament requires only about half this number of notes. Still, if twelve only are used it is possible to modulate to keys closely related to the original, and this satisfied composers of the classical age.

In the eighteenth century another form of temperament began to attract the attention of musicians, particularly of J. S. Bach, in which every note on the instrument bore a constant frequency-ratio to its neighbours. This *equal temperament* allows of all possible modulations since it makes every key equally right—or wrong. It is to this temperament that every pianoforte and organ is nominally tuned to-day.

Much has been written about the relationship of the notes of the musical scale to the harmonic series. While it is true that on an instrument tuned in just, i.e. unadulterated, temperament, the lower harmonics on any fundamental as key-note are to be found as fundamentals of notes higher up the scale, the odd harmonics from the seventh upwards are not so found. Furthermore, in equal temperament, these notes will themselves have been so doctored that coincidence of a harmonic with another fundamental fails save where the order of the harmonic is an integral power of two, i.e. the interval between the two notes in question is a multiple octave (Lloyd, 1943). The overtones do make contact with the practice of tuning at these points, viz. the octaves, because it is from the octaves that the tuner *ranges* his scaling. It so happens that the higher overtones of a struck string of finite thickness are inharmonic, as Rayleigh showed theoretically and as has been proved by experiment on the pianoforte recently (Shuck and Young, 1943). This is due to the increasing importance of the stiffness factor when the string is vibrating in a large number of nodal segments. Thus, if the tuner adheres strictly to equal temperament, the fundamentals of notes high in the treble will be flat to the higher overtones of notes of the same name low

in the bass, and as the ear takes cognisance of such dissonances, the tuner purposely "stretches" the upper octaves to make them correspond with the inharmonic overtone of the lower octaves.

Of course, if one is a purist one may prefer to adopt just temperament for the pianoforte and either play entirely in one key or have the necessary extra notes for playing in as many keys as one wishes. The reed organ on considerations of space or, nowadays, the electronic organ, is best adapted for this monumental construction. Reed instruments with not merely white and black notes, but red, green and yellow ones, which were built with this object in the last century, are to be found in the Science Museum, London ; but whether their owners acquired any facility in playing them is another matter. It has been suggested that an electronic organ should be built with all the oscillators in just temperament in the scale of C major, but that as soon as the player modulates to, say, G major and has established that key, he should press a switch which would make the necessary changes in the frequencies of the whole gamut in the form, presumably, of additional condenser capacities added to or subtracted from those already in the oscillators, to put them all in just intonation with the G's. Thirty-six such stops would be needed to cover all possible modulations in the major, so that the instrument would be quite complicated even at that, but more difficulties would arise in the minor (Williamson, 1944).

What has been written so far applies to Western music. In Eastern music the scale is longer, in the sense that although the octave is still the basis, it is divided into more intervals than the major and minor scales. Thus the scale of modern Egyptian music as played on the qanun, an instrument of plucked strings, has fifteen intervals to the octave (Mosharrafa and Mokhtar, 1939). Rather curiously, considering the difference in origin and evolution, nine of these correspond quite closely to the European scale of semi-tones, but the interval between the third and sixth semi-tones of our scale is divided into four instead of three intervals, while an extra note is interposed between our eleventh and twelfth semi-tones. An investigation of Hindu music on similar lines, i.e. with the use of frequency meters, is also promised (Sitaraman, 1944).

The case of obsolete musical scales is rather different. There one must rely on descriptions, often obscure and sometimes in cypher, in the writings of the past. A high place has been claimed for ancient Greece in the development of musical theory, since, with Pythagoras, arithmetical science was first applied to the determination of the lengths of strings and pipes. Some modern reconstructions of the theoretical Greek modes have been made (notably by the late Dr. Perrott, who at University College, London, had a beautiful set of Tinsley tuning forks on resonators with which he used to demonstrate ancient Greek music to students), but it is very doubtful whether, considering the primitive methods of constructing musical instruments at that time, anyone actually heard a Greek scale before the nineteenth century.

Indeed, the limitations of practical music-making will always act as a brake on applied acoustical theory. It must also be remembered that the appeal of music is to the ear, and in so far as physical theory assumes an ideal sound-receiver, which the ear is not, physics without physiology and psychology must be incapable of explaining completely the basis of music. The limitations of

"apparatus" are usually accepted by the composer at the age in which he writes, and the ear, as final arbiter, cannot be argued with. One is thinking in this connection of the sort of limitation whereby on a musical instrument the frequency of a tone changes with the energy put into it, and whereby the pitch of a tone of constant frequency appears to the ear to change with intensity. The subjective addition in the ear of a missing fundamental to a harmonic series sounded simultaneously is another instance.

REFERENCES

BÄR, R., 1940. *Helv. phys. Acta*, **13**, 61.
BEHR, A., 1943 a. *Mech. World*, **113**, 191.
BEHR, A., 1943 b. *Metal Ind.*, **63**, 422.
BERANEK, L. L., 1942. *J. Acoust. Soc. Amer.*, **13**, 248.
BONER, C. P., 1942. *J. Acoust. Soc. Amer.*, **13**, 244.
CONTENT, E. J. and GREEN, L., 1944. *Proc. Inst. Radio Engrs.*, **32**, 72.
DUFTON, A. F., 1941. *Phil. Mag.*, **32**, 259.
DUFTON, A. F., 1942. *Phil. Mag.*, **33**, 151.
FLETCHER, H., 1941. *J. Acoust. Soc. Amer.*, **13**, 89.
HILL, S., 1942. *J. Inst. Elect. Engrs.*, **89** (iii), 124.
LLOYD, LL. S., 1943. *Phil. Mag.*, **34**, 472, 624, 674.
LUDLOFF, H. F., 1940. *J. Acoust. Soc. Amer.*, **12**, 193.
MAA, D. Y., 1941. *J. Acoust. Soc. Amer.*, **13**, 170.
MASON, C. A. and MOIR, J., 1941. *J. Inst. Elect. Engrs.*, **88** (iii), 175.
MEYER, E. and BUCHMANN, G., 1938. *Akust. Z.*, **3**, 132.
MOSHARRAFA, A. M. and MOKHTAR, M., 1939. *Proc. Phys.-Math. Soc., Egypt*, **1**, 9.
POHLMAN, R., 1939. *Z. Phys.*, **113**, 697.
SABINE, P. E., 1942. *J. Acoust. Soc. Amer.*, **14**, 74.
SCHMIDT, G. and EHRET, L., 1937. *Z. Elektrochem.*, **43**, 869.
SCHMIDT, G. and ROLL, A., 1939. *Z. Elektrochem.*, **45**, 769.
SHUCK, O. H. and YOUNG, R. W., 1943. *J. Acoust. Soc. Amer.*, **15**, 1.
SIEGEL, S., 1940. *J. Chem. Phys.*, **8**, 860.
SIEGEL, S., 1944. *J. Acoust. Soc. Amer.*, **16**, 26.
SIEGEL, S., and QUIMBY, S. L., 1936. *Phys. Rev.*, **49**, 663.
SITARAMAN, M. L., 1944. *J. Annamalai Univ.*, **11**, 178.
SOKOLOV, S., 1935 a. *Z. tech. Phys. U.S.S.R.*, **2**, 522.
SOKOLOV, S., 1935 b. *Acta Physiochem.*, **3**, 939.
TROST, A., 1943. *Z. Ver. dtsch. Ing.*, **87**, 352.
VOLCKMANN, J. E., 1942. *J. Acoust. Soc. Amer.*, **13**, 234.
WHITE, W. B. L., 1943. *J. Acoust. Soc. Amer.*, **15**, 12.
WILLIAMSON, C., 1944. *J. Acoust. Soc. Amer.*, **15**, 173.
ZWIKKER, C. J. and KOSTEN, C. W., 1941. *Physica*, **8**, 469.

PHASE SEPARATION IN AQUEOUS COLLOIDAL SYSTEMS

By P. KOETS,
Stanford University, California

§1. INTRODUCTION

THE greater part of the literature on the formation of liquid or semi-liquid phases in aqueous colloid dispersions was, until recently, concerned with the application of the Gibbs' Phase Rule to these phenomena. The validity of such application has been denied on the grounds that the energy located in the interfacial surfaces cannot be neglected, as is required by Gibbs' deduction of the phase rule, and that the thermodynamic potential of the phases is not independent of form, surface condition and other properties of these phases (Galeotti, 1906). According to other critics, a colloidal dispersion cannot be regarded as a homogeneous one-phase system since the dispersed colloid, though consisting of particles of ultramicroscopic dimensions, forms, in fact, a separate phase (Büchner, 1926 ; von Buzagh, 1936). Against this it has been argued that the heterogeneity of a colloid dispersion is only relative, and that it can be regarded as homogeneous, in the sense of the phase rule, provided the volume element selected is large compared with the colloid micelle, in the same way that the volume studied in molecular solutions is large with respect to the molecules of the solution (Büchner, 1926 ; McBain, Vold, R. D., and Vold, M. J., 1938).

Experimental observations have been made to test whether number, composition, and other parameters of the phases separating from colloidal solutions follow the requirements of the phase rule. This has been denied in the case of protein systems, e.g. gelatin and sulphosalicylic acid (Ostwald and Köhler, 1927), but affirmed in extensive studies of solutions of typical colloidal electrolytes, e.g. soap systems (McBain, Vold, R. D., and Vold, M. J., 1938).

In these discussions the mechanism of the formation of such new phases and the nature of the inter-molecular (intermicellar) forces involved were not considered. Indeed, the phase rule applies only to the conditions under which phases separate and to the resulting final state of equilibrium, and is not concerned with the processes which lead to this equilibrium.

Recently a number of new instances of phase separation in colloid systems has been observed. The importance of these phenomena for biology and for application in many other fields has stimulated their further study, which has centred round the mechanism of the formation of the phases rather than the application of the phase rule. The study of these processes has contributed to the further knowledge of colloid systems and of the forces acting between colloid particles and has in turn led to the uncovering of examples of liquid phase separation in molecularly dispersed electrolyte solutions.

Phase separation in colloid dispersions can be classified roughly as occurring :

(i) in dilute solutions of hydrophilic colloids containing particles of opposite electric charge ;

(ii) in highly concentrated solutions of hydrophilic colloids of the same electric charge;

(iii) in colloid solutions containing particles of anisotropic dimensions.

In (i) and (iii) the newly formed phase has a higher concentration of colloid component than the original stable dispersion. The particles in this phase are at distances varying from 100 to 1000 A. from each other. In order that their structure can be understood, the question to be answered here concerns the nature of the forces between the colloid particles at such distances.

§2. SURFACE CONDITIONS IN STABLE COLLOIDAL DISPERSIONS

A substance displays colloid properties in dispersion when it is subdivided in a medium to particles of a size intermediate between those present in common molecular dispersions and in coarse suspensions (conventionally ranging from approximately 1 to 100 millimicrons). The particles may be polymolecular aggregates (Au, AgI) or, as in the case of bio-colloids (proteins, carbohydrates), the single molecule may be of a size which places it within the realm of colloid dimensions.

The state of subdivision is not in itself sufficient to ensure stability of the dispersion. On the contrary, particles of this size are subject to intense Brownian motion and would rapidly adhere together on collision unless conditions in the surface are such as to prevent effective material contact. It was recognized relatively early in the development of colloid chemistry that one of these conditions can be an electrical one (Hardy). In stable colloid dispersion the micelles can carry an electric charge, which can be due to preferential adsorption of an electrolyte ion from the solution or, notably in colloid dispersions of bio-chemical material, can result from the dissociation of surface groups which are an integral part of the constituent molecule. In either case the charged micelle is surrounded by an equivalent amount of oppositely charged ions. The distribution of these counter-ions in the atmosphere around the particle gives rise to a potential gradient (Helmholtz). For as far as this potential gradient is situated outside the adsorbed water layer which moves with the particle (electrokinetic potential), it is connected with the stability of the colloid dispersion, and its lowering by the addition of electrolytes beyond a certain critical value may result in destruction of the stability. The amount of electrolyte necessary to effect such a decrease is in general less the higher the valency of the added ion of charge opposite to that of the micelle (Hardy-Schulze). This influence on the electrokinetic potential, and consequently on the colloid stability, is a complicated one since polarization and adsorption influences are superimposed on the electrostatic relations between ions and colloid surface.

These electric surface conditions can be shown to occur in all stable aqueous colloid dispersions. In dispersions of hydrophilic colloids, however, lowering of the electrokinetic potential does not lead necessarily to destruction of the stability, since hydration of the hydrophilic particle can by itself act as a stabilizing condition. Hydration is generally assumed to result from orientation of water dipoles in electric fields emanating from the particle surface. The orientation is assumed to continue for some distance from the particle and to decrease gradually as thermal agitation overcomes the energy of orientation. More recently,

hydration has been attributed to hydrogen bonding between suitable groups in the surface and water molecules; it is further assumed that this bonding between the water molecules continues for some distance into the surrounding solution. It has been shown that the properties of this water of hydration differ from those of free water in that its density is higher and its solvent power lower.

No exact data are available regarding the interdependence of electric charge and hydration, the state and distribution of the counter-ions in the oriented water layers and their behaviour during electrophoresis and electro-osmosis. In general our knowledge of colloid hydration does not go far beyond the qualitative state.

§3. COMPLEX COACERVATION*

The separation of a liquid phase from aqueous colloid dispersions has been observed in many instances when dispersions of two hydrophilic colloids with micelles of opposite electric charge are mixed. The phase separates in the form of droplets of microscopic dimensions which cause the mixture of the sols to assume a turbid appearance. The droplets coalesce with greater or less rapidity, depending on their viscosity, to form a liquid layer as a macro-phase.

On the basis of the conclusions regarding surface conditions of a hydrophilic colloid particle in stable dispersion the following explanation has been offered for the formation of the liquid phase. The micelles of opposite polarity attract each other by Coulomb forces after mutual neutralization of the ionic atmospheres. The attraction results in an approach of the particles, and the outer, weakly oriented water dipoles are removed from the oriented shells. This approach will continue until the force of attraction is balanced by the resistance of the remaining, more rigidly oriented, water. The droplet is then a conglomeration of oppositely charged particles separated individually by rigidly oriented water. (*Acervus*, a swarm ; coacervation, a swarming together.)

This conception of the structure of the liquid phase is in good agreement with its properties. The coacervate phase is formed over a certain, usually rather wide, range of proportions of the two oppositely charged colloids. The two components are present in the liquid complex in approximately the ratio in which they are mixed in solution. At one proportion the resulting droplets display no outward electric charge, that is they do not move in an electric field. The electric charges carried by the masses of the two colloids present in the complex at that ratio are equivalent. On either side of this point the droplets show anodic or cathodic migration, depending on the sign of the component which preponderates. The amount of water in the coacervate is found to be a minimum at the iso-electric proportion where the charges balance each other exactly and the approach is consequently closest. On either side of this point the amount of water retained between the particles increases as the excess charge of that component which is in excess causes an extending tendency in the complex. When one of the colloids is present in large excess no visible separation takes place. When the droplets are placed in a sufficiently strong electric field, electro-disintegration can be observed as the individual micelles leave the complex in opposite directions.

* See Kruyt and Bungenberg de Jong, 1930 ; Bungenberg de Jong and collaborators, 1929–36, 1932–39, 1934–35, 1935–39. For summaries, see Bungenberg de Jong, 1936–37 b. (For Bungenberg de Jong's papers please refer to de Jong, H. G. Bungenberg.—Editor.)

The conclusion that formation of the liquid phase is accompanied by a certain degree of dehydration of the participating particles is confirmed by viscosity measurements. When colloid solutions, containing particles which do not interact, are mixed, the viscosity of the mixture can be calculated by the rule of additivity from the viscosities of the components. When coacervation takes place between colloids of opposite charge, the viscosity of the mixtures is less than that calculated from additivity. The deviation from the calculated value is greatest at the iso-electric proportion.

It has been suggested that "the very intense fields which develop when the particles come close together are apparently able to draw water in between the particles and so hold them apart", and that "the field rises to a value so high as to cause *increased* hydration" (Langmuir, 1938). This point of view does not seem to agree so well with the experimental facts as the one previously described, namely, that the water of hydration is originally oriented round the hydrophilic particles and is *forced out* to the point where the remaining hydration balances the electrostatic attraction. When dispersions of two *hydrophobic* colloids of opposite charge are mixed (e.g. iron hydroxide and arsenic trisulphide) complex formation is also observed over a range of proportions, but no liquid phase is observed under any circumstance. If water were drawn into the field between the particles, liquid separation might be expected in the case of hydrophobic colloids as well, unless it could be shown that the intensity of the field is of an entirely different order in the latter case. Viscosity measurements also indicate a definite dehydration of the particles during the formation of the coacervate phase, which means that the water of hydration is present before the complex is formed, and is not drawn in when a field between the particles is set up. The difference between these two points of view would seem trivial were it not that these studies may throw further light on the question of the extent of water binding round hydrophilic particles. It cannot be denied, however, that though the hydration of the particles in the coacervate state is less than that of free micelles, the binding of the remaining water is stronger since, in the coarcervate, it is subjected to an electrostatic field superimposed on the field of orientation. No mathematical treatment of the energy relations involved has as yet been attempted.

Since the liquid phase owes its structure to a balance of forces, it is highly sensitive to changes in the solution with which it is in contact. Any change which influences either of the forces will of necessity bring about a change in internal structure, composition, volume and other properties of the coacervate phase. The addition of neutral electrolytes results in a lowering of the Coulomb attraction by shielding of the electric charges on the particles and consequently leads first to an increased water content in the coacervate and ultimately to its re-dispersal. The quantitative aspect of the influence of neutral electrolytes is expressed by the Hardy-Schulze rule for the lowering of the electrokinetic potential. Since, on addition of a neutral electrolyte, positive and negative micelles are affected simultaneously, a Hardy-Schulze rule is observed for both anions and cations with regard to the amount of electrolyte necessary to bring about re-dispersal of the coacervate or to inhibit its formation.

The amount of a particular electrolyte, e.g. KCl, required to bring about re-dispersal depends upon the nature of the individual colloid components in the

coacervate and, on this basis, strong and weak coacervators have been distinguished. Among the strong coacervators are substances which display a high density of charge on the micelle in dispersion, such as the nucleinates, which are important building stones in vital biological structures. Weak coacervators, which usually have a low density of charge in solution, include such substances as starch and glycogen, which (as storage material) play a latent rôle in biology.

The extent to which a coacervate phase will display liquid character depends on the nature of the components and on the electrolyte content of the surrounding solution.

§4. AUTOCOMPLEX COACERVATION

Similar phenomena to those described above for solutions of two hydrophilic colloids can be observed when one of these colloids is replaced by a suitable ion which itself has no colloid properties. The ion must then carry a charge opposite to that of the one remaining colloid and must in general be polyvalent or complex multipolar. Thorium chloride will cause liquid coacervate droplets to separate from a solution of a negative hydrophilic colloid, e.g. gum arabic, over a certain range of thorium concentrations, the electric charge of the droplets changing with increasing thorium concentration from negative, through an iso-electric condition, to positive. The coacervate re-disperses at higher thorium concentrations.

It was assumed originally that the polyvalent ion is absorbed on the colloid particle, neutralizing a monovalent centre of charge and thus leading to a mosaic of charges on the surface of the micelle. Later evidence, including the observation of electro-disintegration of these autocomplex coacervates and the occurrence of liquid phase separation in inorganic salt solutions, led to the conclusion that the polyvalent, non-colloidal ion occupies a position wholly equivalent to that of the colloid component. The relationship between colloid micelle and polyvalent ion in the complex is thus the same as that described for colloid micelles of opposite polarity : a Coulomb attraction balanced by the hydration of hydrophilic micelle and ion.

It is found as a general rule that the minimum valency of the ion which will bring about autocomplex coacervation in a solution of a given colloid is less the higher the density of that colloid in dispersion. Nucleinates can be brought to coacervation by divalent cations, gum arabic requires an ion which is at least quadrivalent, and amylose in aqueous solution shows no coacervation even with a hexavalent cation. When the valency of the opposing ion is too low to lead to an effective attraction, visible coacervation can usually be realized by the use of a suitable dehydrating agent which will shift the balance in favour of the attractive tendency.

§5. THE EFFECT OF VAN DER WAALS FORCES IN AUTOCOMPLEX COACERVATION

Among the substances which can be separated from their stable colloid dispersions in the form of liquid autocomplex coacervates are the biologically important phospholipids, e.g. lecithin. When purified, i.e. freed in particular of accompanying fats and sterols, lecithin requires a cation which is at least quadrivalent in order to establish sufficient opposition in polarity. The addition of a dehydrating agent is necessary for liquid separation in the case of a commonly occurring

divalent ion such as calcium. It was observed, however, that, in the case of phospholipids, coacervation with calcium ions can also be obtained when certain organic compounds such as fatty acids, higher alcohols, sterols, and so forth, which are not conventional dehydrating agents, are incorporated in the system.

In the lecithin molecule, groups of different functional character can be distinguished. One end of the molecule is formed by the combination of the hydrophilic phosphoric acid and choline group. This part of the molecule is amphoteric and displays a weak negative charge under physiological conditions. The other end of the lecithin molecule is formed by the hydrocarbon chains of the esterified fatty acids and is entirely hydrophobic in character. The autocomplex relationship between lecithin and polyvalent ions involves exclusively the hydrophilic part of the molecule. The attraction between these groups and divalent calcium ion is evidently too weak to overcome the repelling tendency of the hydration. Organic substances which are predominantly hydrophobic in nature can now exert sufficient additional attraction through van der Waals forces on the hydrocarbon chains of the lecithin molecule to lead to separation of a coacervate phase. The structure of this phase is then determined by three forces : an attractive force of electrostatic nature, an attractive van der Waals force and a repellent force of hydration. Since the action of the attractive forces is exerted on different ends of the molecule, a certain orientation may result in the complex, and this is reflected by morphological aspects of these structures (e.g. oriented bimolecular films).

The influence of neutral electrolytes on the coacervate relationship as far as it concerns the hydrophilic part of the complex is the same as that described previously for complex coacervation. Thus added sodium ion has an antagonistic action on the binding force of calcium.

The amount of water retained between the lecithin molecules in these coacervates depends not only on the effective electric antagonism of calcium ion and hydrophilic negative groups but is now also determined by the contracting action of the added hydrophobic organic substance. By keeping the calcium-ion concentration of the system constant it is then possible to compare the van der Waals attraction exerted by different organic substances on the same hydrocarbon chains by measuring the volumes (as expressing water content) of the coacervates separating in each instance. The influence of members of a number of homologous series has been investigated in this way, partly in lecithin systems and partly in oleate systems where conditions are analogous.

It has been suggested that in the protoplasmic membrane of the animal cell, phospholipid, cholesterol, calcium and sodium ions are held together in such a relationship as a coacervate film and that the permeability of the membrane is determined, at least in part, by the distance between the phospholipid molecules and the amount of water retained in the complex, which in turn depend on cholesterol and calcium concentration (as well as on *p*H, electrolyte concentration of the surrounding medium, as in all coacervates) (Bungenberg de Jong and Bonner, 1935).

In studies of the relative effectiveness of the attraction exerted by aromatic hydrocarbons on oleates it was found (i) that this function increases, as might be expected, in the order benzene—naphthalene—anthracene : (ii) that, however,

the attraction is greater in the case of hydrocarbons with phenanthrene structure, increasing in the series pyrene—phenanthrene—cholesterol (Koets and Bungenberg de Jong, 1938). The importance of biological compounds which have phenanthrene structure, such as sex hormones, vitamin D, bile-acid derivatives, carcinogenic hydrocarbons, and so forth, becomes apparent in this respect. The replacement of cholesterol in a protoplasmic membrane of coacervate structure by one of these substances, or by a modification of its own molecule, leads then primarily to a change in the intermolecular distances of the membrane molecules and consequently to an altered water content and permeability.

§6. LIQUID-PHASE SEPARATION IN SALT SOLUTIONS

In the theory offered for the explanation of complex coacervation in colloid systems (see §3) the colloidal dimensions of the particles do not enter specifically. It was consequently anticipated that under proper conditions such a liquid-phase separation might occur in any system containing units of opposite polarity and a sufficient degree of hydration, whether these units are in colloidal or in true solution. The conditions should be such that rigid orientation of water dipoles between the components is favoured. This may be expected to occur at low temperatures with decreased kinetic energy of the water dipoles. Since as a rule the state of the liquid phase will be intermediate between homogeneous solution and solid crystal, the conditions should be similar to those favouring super-saturated solutions, so that high concentrations should be used. From Haber's conclusions about the relative velocity with which dipoles and multipoles pass from an un-ordered state into a crystal lattice it followed further that the ions involved should be preferably either polyvalent or complex multipolar.

A number of examples of the formation of liquid phases in salt solutions when the proper conditions were used, were thus realized with such cation components as barium, strontium, tin, cadmium, lead and complex cobalt ions with anions consisting of molybdic acid, complex cobalt salts and numerous carboxylic acids (Kruyt and Bungenberg de Jong, 1935 ; Bungenberg de Jong and Teunissen-van Zijp, 1936). In the majority of cases realized the liquid state was not the stable one and crystallization to an "insoluble salt" occurred. However, in such combinations as novocaine bichromate, examples were found where the liquid phase could be shown to be the stable one above certain temperatures (Bungenberg de Jong and Holleman, 1937).

§7. PHASE SEPARATION IN COLLOID SOLUTIONS CONTAINING PARTICLES OF ANISOTROPIC DIMENSIONS

Spontaneous separation of liquid phases has been observed in dispersions of colloid material in the apparent absence of an opposition of charges on the colloid particles, or between them and a multivalent ion. This occurs in hydrous oxide sols where separate and often bi-refringent liquid layers have been designated by such terms as Schiller layers (in the case of ferric oxide and tungstic oxide sols) or as tactoids (in sols of vanadium pentoxide) (Zocher, 1925 ; Zocher and Jacobsohn, 1927, 29 ; Bergmann, Beer and Zocher, 1938). A more recent example is the separation of a liquid phase from solutions of tobacco mosaic virus (a protein) when the concentration exceeds a certain limit (Bawden, Pirie, Bernal and Fankuchen, 1936 ; Bawden and Pirie, 1937 ; Freundlich, 1937 ; Lauffer and

Stanley, 1944). Distances between particles in these phases may vary from 125 A. to as high as 8000 A. Though they differ greatly in chemical composition, the substances in these instances have the property in common that their micelles have distinctly anisotropic dimensions, rod-shaped in the virus and vanadium pentoxide particles, disc-shaped in those of ferric hydroxide.

The nature of the forces involved in the formation of these conglomerates is still being debated, though there seems to be general agreement that here too the final equilibrium is the result of a balance of forces. As the repulsive force, the electrostatic repulsion between the similar charges of the micelles is usually assumed. As the attractive force, several authors have suggested van der Waals forces, which must then, however, be active over far greater distances than had previously been accepted (Hamaker, 1937 a, 37 b ; Derjaguin, 1940). The subject has been treated diagrammatically by depicting suitable energy graphs for the change of these forces with the distance between the micelles. In this way it is of course possible to construct diagrams with an energy minimum at any desired distance between the particles. As long, however, as the nature of the forces is unknown, and the data are lacking to compute such graphs, diagrams of this kind would seem to have little more than hypothetical value.

It has been shown subsequently that the assumption of long-range van der Waals forces becomes unnecessary when the interaction of the particles and their counter-ions as a whole is considered (Langmuir, 1938). The "lattice" Coulomb forces in the system supply the attractive force which by itself is sufficient to overcome the dispersive tendency due to thermal agitation. In order to attain an equilibrium with the particles at a distance from each other, a balancing repulsive force is now required which comes into being as, with decreasing distance between the particles, their charge increases and, with it, the electrostatic repulsion.

So far the hydration of this type of colloid has not been considered in the literature in explaining the formation of the liquid phase, though the virus protein is a typical hydrophilic colloid and the metal hydroxide micelles have strongly hydrophilic surfaces. In the elongated structure of this type of micelle, the possibility exists of an alternating arrangement of positive and negative charges in the protein (even though the overall charge is negative) and of dipoles in the hydroxides. That such an arrangement can lead to rigid structures of oriented water has been suggested previously in the case of proteins (de Boer and Dippel, 1933) and of carbohydrates (Koets and Kruyt, 1938).

REFERENCES

Bawden, F. C., Pirie, N. W., Bernal, J. D. and Fankuchen, I., 1936. *Nature, Lond.*, **138**, 1051.

Bawden, F. C. and Pirie, N. W., 1937. *Proc. Roy. Soc.*, B, **123**, 274.

Bergmann, P., Beer, P. Löw and Zocher, H., 1938. *Z. phys. Chem.*, A, **181**, 301.

de Boer, J. H. and Dippel, C. J., 1933. *Rec. Trav. chim. Pays-Bas*, **52**, 214.

Buchner, E. H., 1926. *Colloids as One Phase Systems* in J. Alexander's *Colloid Chemistry* (New York : Chemical Catalog Co.), **1**, 126.

Buzágh, A. von, 1936. *Kolloidik* (Dresden : Steinkopff).

Derjaguin, B., 1940. *Trans. Faraday Soc.*, **36**, 203.

Freundlich, H., 1937. *J. phys. Chem.*, **41**, 1151.

Galeotti, G., 1506. *Z. phys. Chem.*, **54**, 727.

Hamaker, H. C., 1937 a. *Rec. Trav. chim. Pays-Bas*, **56**, 1 and 727.

Hamaker, H. C., 1937 b. *Physica*, **4**, 1058.

DE JONG, H. G. BUNGENBERG, 1936. *La Coacervation et son Importance en Biologie* (Paris : Hermann et Cie).
DE JONG, H. G. BUNGENBERG, 1937 a. *Kolloidzschr.*, **79**, 223 and 334.
DE JONG, H. G. BUNGENBERG, 1937 b. *Kolloidzschr.*, **80**, 221 and 350.
DE JONG, H. G. BUNGENBERG and BONNER, J., 1935. *Protoplasma*, **24**, 198.
DE JONG, H. G. BUNGENBERG and COLLABORATORS, 1929–36. *Biochem. Z.*, **212–288**.*
DE JONG, H. G. BUNGENBERG and COLLABORATORS, 1932–39. *Protoplasma*, **15–33**.*
DE JONG, H. G. BUNGENBERG, and COLLABORATORS, 1934–35. *Rec. Trav. chim. Pays-Bas*, **53–54**.*
DE JONG, H. G. BUNGENBERG and COLLABORATORS, 1935–39. *Proc. Acad. Sci. Amst.*, **38–42**.*
DE JONG, H. G. BUNGENBERG and HOLLEMAN, L. W. J., 1937. *Proc. Acad. Sci. Amst.*, **40**, 69.
DE JONG, H. G. BUNGENBERG and TEUNISSEN-VAN ZIJP, L., 1936. *Proc. Acad. Sci. Amst.*, **39**, 1103.
KOETS, P. and DE JONG, H. G. BUNGENBERG, 1938. *Protoplasma*, **30**, 206.
KOETS, P. and KRUYT, H. R., 1938. *Kolloidzschr.*, **82**, 315.
KRUYT, H. R. and DE JONG, H. G. BUNGENBERG, 1930. *Kolloidzschr.*, **50**, 39.
KRUYT, H. R. and DE JONG, H. G. BUNGENBERG, 1935. *Proc. Acad. Sci. Amst.*, **38**, 714.
LANGMUIR, I., 1938. *J. Chem. Phys. Phys.*, **6**, 873.
LAUFFER, M. A. and STANLEY, W. M., 1944. *The Colloid Chemistry of Purified Viruses* in J. ALEXANDER'S *Colloid Chemistry* (New York: Reinhold Publishing Co.), **5**, 785.
MCBAIN, J. W., VOLD, R. D. and VOLD, M. J., 1938. *J. Amer. Chem. Soc.*, **60**, 1866.
OSTWALD, W. and KOHLER, R., 1927. *Kolloidzschr.*, **43**, 131 and 151.
ZOCHER, H., 1925. *Z. anorg. Chem.*, **147**, 91.
ZOCHER, H. and JACOBSOHN, K., 1927. *Kolloidzschr.*, **41**, 220.
ZOCHER, H. and JACOBSOHN, K., 1929. *Kolloid-Beihefte*, **28**, 167.

PAPERS BY H. G. BUNGENBERG DE JONG AND COLLABORATORS

KRUYT, H. R. and DE JONG, H. G. BUNGENBERG, 1930. *Kolloidzschr.*, **50**, 39. First general article.

Series in Biochemische Zeitschrift †

H. G. BUNGENBERG DE JONG with :

DEKKER, W. A. L., 1929. (**212**, 318.) System gum arabic and gelatin : viscosity measurements.

GWAN, ONG SIAM, 1930 a. (**221**, 166.) General observations on the viscosity of sol mixtures.

GWAN, ONG SIAM, 1930 b. (**221**, 182.) System serum albumen and gum arabic : viscosity measurements.

DEKKER, W. A. L. and GWAN, ONG SIAM, 1930. (**221**, 392.) Complex coacervation under physiological conditions, systems clupein and nucleic acid : measurements of salt resistance.

DEKKER, W. A. L., 1930. (**221**, 403.) Behaviour of coacervates in an electric field : electrophoresis, electro-disintegration.

KAAS, A. J. W., 1931. (**232**, 338.) Electrophoretic behaviour of droplets of organic fluids incorporated in coacervate drops.

WESTERKAMP, R. F., 1931. (**234**, 367.) Complex coacervation of lecithin (pos. or neg.) with other colloid components : salt resistance, morphological structure.

LENS, J., 1931. (**235**, 174.) Equivalent weight and reversal of charge of hydrophilic colloids.

LENS, J., 1931. (**235**, 174.) Autocomplex coacervation : gum arabic and hexol salt, salt resistance, etc.

WINKLER, K. C., 1932. (**248**, 115.) Autocomplex coacervation : gum arabic and lower-valent ions, influence of dehydrating non-electrolytes (acetone).

WESTERKAMP, R. F., 1932. (**248**, 131.) Autocomplex coacervation of lecithin and lower-valent ions : influence of water soluble dehydrating agents (alcohol, resorcinol).

* A complete list of the papers by Bungenberg de Jong and his collaborators, with an indication of their contents, is appended.

† In chronological order.

WESTERKAMP, R. F., 1932 a. (**248**, 309.) Autocomplex coacervation of lecithin and lower-valent ions : influence of triolein, oleic acid, cholesterol.

WESTERKAMP, R. F., 1932 b. (**248**, 335.) Lecithin autocomplex coacervates : behaviour in electric field, morphological appearance, bearing on membrane permeability.

LENS, J., 1932. (**254**, 15.) Coacervation of gum arabic and trypaflavin : gelatin and sodium picrate and tropaeolin O.

MENALDA, F. A., 1933. (**254**, 35.) Autocomplex coacervation of trypaflavin and sodium pyrophosphate—and K-methanetrisulphonate.

MENALDA, F. A., 1933. (**257**, 62.) Autocomplex coacervation of nucleic acid with hexol salt and with lower-valent ions, including Ca.

WINKLER, K. C., 1933. (**259**, 436.) Further details on the autocomplex coacervation of gum arabic and hexol : influence of the sol concentration.

DE JONG, H. G. BUNGENBERG, 1933. (**259**, 442.) Theoretical considerations on the mechanism of the binding of polyvalent ions.

H. G. BUNGENBERG DE JONG with :

VAN DER HORST, J. H. and LAFLEUR, A., 1933. (**260**, 161.) Specific influences in the autocomplex coacervation of gum arabic and hexol, specific alcohols and salts.

VAN DER LINDE, P., 1933 a. (**262**, 161.) Equivalent weight of arabinates (H-, Na-, NH_4-, Sr-, luteo-, hexol-) and considerations on the origin of the electric charge.

VAN DER LINDE, P., 1933 b. (**262**, 390.) (*a*) Further examples of autocomplex coacervation : agar, chondroitin sulphuric acid, carrhagen, *semen lini*, positive proteins and polyvalent ions. (*b*) Correlations of the reciprocal hexol number and the tendency to form coacervates.

DE HAAN, A., 1933. (**263**, 33.) (*a*) Further correlations of reciprocal hexol number, ash content, salt resistance, etc. (*b*) General remarks on co-existent coacervates.

SAUBERT, G. G. P., 1936. (**288**, 1.) Complex flocculation and coacervation of the type : colloid amphoteric ion—colloid anion—crystalloid cation.

Series in Protoplasma †

DE JONG, H. G. BUNGENBERG, 1932. (**15**, 110.) Coacervation and its importance in biology. General summary of complex coacervation.

H. G. BUNGENBERG DE JONG with :

BONNER, J., 1935. (**24**, 198.) Lecithin autocomplex coacervates : formation of bimolecular films, bearing on protoplasmic membranes and permeability.

BOOY, H. L., 1935. (**24**, 319.) Influence of combinations of three salts (NaCl, $CaCl_2$, $MgCl_2$) on autocomplex coacervates of phospholipids.

SAUBERT, G. G. P., 1936. (**26**, 282.) Antagonism of neutral electrolytes in the reversal of charge of adsorbed arabinate and lecithin films.

WESTERKAMP, R. F., 1936. (**27**, 32.) Electromotoric force in the chain : 0·1 NKCl—coacervate—1·0 NKCl.

SAUBERT, G. G. P., 1937 a. (**28**, 329.) Influence of vapours of organic substances on lecithin and oleate coacervates.

SAUBERT, G. G. P., 1937 b. (**28**, 352.) Participation of cholesterol, phosphatidic acid and proteins in lecithin coacervates.

SAUBERT, G. G. P., 1937 c. (**28**, 498.) Qualitative measurements of the influence of hydrocarbons on oleate coacervates.

BOOY, H. L. and SAUBERT, G. G. P., 1937. (**28**, 543.) Quantitative measurements of the influence of aliphatic alcohols on oleate and on lecithin coacervates.

SAUBERT, G. G. P. and BOOY, H. L., 1938 a. (**29**, 481.) Influence of temperature changes on oleate coacervates.

SAUBERT, G. G. P. and BOOY, H. L., 1938 b. (**30**, 1.) Correlation between constitution and action of different classes of organic substances (except amines) on oleate coacervates.

BOOY, H. L. and SAUBERT, G. G. P., 1938 a. (**29**, 536.) Influence of *p*H, fatty acids and CO_2 on oleate coacervates.

† In chronological order.

BOOY, H. L. and SAUBERT, G. G. P., 1938 b. (**30**, 53.) Influence of amines on oleate coacervates : rôle of nitrogen.

KOETS, P., 1938. (**30**, 206.) Quantitative measurements of the influence of aromatic and aliphatic hydrocarbons on oleate coacervates.

DE HAAN, A. and WAKKIE, J. G., 1938. (**30**, 592.) Ion action and permeability : electrophoresis measurements in plasmolysed protoplasts and lecithin coacervates.

HARTKAMP, J. L. L. F., 1939. (**31**, 550.) Formation of hyaline vesicles at the surface of paramecium : discussion of the plasma membrane.

BANK, O., 1939 a. (**32**, 489.) Interaction of biocolloids and dyes.

BANK, O., 1939 b. (**33**, 321.) Morphology of complex gels.

Series in Recueil des Travaux chimiques des Pays-Bas et de la Belgique †

DE JONG, H. G. BUNGENBERG, 1934. (**53**, 163.) Complex relationships in lyophilic colloidal systems (General classification and review).

H. G. BUNGENBERG DE JONG with :

SENGERS, H. J. C., 1934. (**53**, 171.) Coacervation phenomena in gels of gelatin and gum arabic and other examples : influence of *p*H, neutral salts.

DEKKER, W. A. L., and WINKLER, K. C., 1934. (**53**, 607.) (*a*) Examples of complex and auto-complex flocculi ; (*b*) The salt solubility of globulins at the iso-electric point.

HARTKAMP, J. L. L. F., 1934. (**53**, 622.) Influence of neutral electrolytes on the electrophoretic velocity of suspended coacervates drops.

VAN DER LINDE, P., 1934 a. (**53**, 737.) Autocomplex flocculi of carrhagen mucilage sol.

VAN DER LINDE, P., 1934 b. (**53**, 747.) Autocomplex flocculi of *semen lini* mucilage sol : examples of autocomplex fibrils.

DEKKER, W. A. L. and VAN DER LINDE, P., 1935. (**54**, 1.) Complex and autocomplex sols (one component in excess).

VAN DER LINDE, P. and DE HAAN, A., 1935. (**54**, 17.) Conclusion ; Essential and non-essential specific characteristics of colloidal components of importance in complex relationships : reciprocal hexol number and solvation.

Series in Proceedings of the Royal Academy of Sciences, Amsterdam †

H. G. BUNGENBERG DE JONG with :

VAN DER LINDE, P., 1935 a. (**38**, 419.) Coacervate sols in relation to the theory of colloid stability.

VAN DER LINDE, P., 1935 b. (**38**, 426.) Oriented coacervates and their bearing on the formation of colloid-crystals.

BONNER, J., 1935. (**38**, 797.) Phosphatide autocomplex coacervates as ionic systems and their relation to the protoplasmic membrane.

KRUYT, H. R., 1935. (**38**, 714.) Extension of the theory of complex coacervation to ionical disperse systems : coacervation in solutions of inorganic electrolytes.

BONNER, J., 1935. (**38**, 346.) Coacervation of pectic and pectinic acids.

KOETS, P., 1935. (**38**, 63.) Complex coacervation of amylophosphoric acid and proteins and its possible bearing on the problem of amylopectin.

H. G. BUNGENBERG DE JONG with :

TEUNISSEN-VAN ZIJP, L., 1936. (**39**, 1103.) Further examples of coacervation in inorganic salt solutions.

HOLLEMAN, L. W. J., 1937. (**40**, 69.) Examples of stable unmixing in binary systems : salt-water.

SAUBERT, G. G. P., 1937. (**40**, 295.) Lecithin coacervates and their relation to the protoplasmic membrane.

SAUBERT, G. G. P., 1937. (**40**, 302.) Models for the stimulation of the organs of smell : reaction of organic vapours on complex coacervates.

DE JONG, H. G. BUNGENBERG, 1938 a. (**41**, 643.) Microscopic studies of biocolloid complexes suspended in aqueous medium.—I. Pulsating vacuoles in coacervate drops.

DE JONG, H. G. BUNGENBERG, 1938 b. (**41**, 646.) Microscopic studies of biocolloids suspended in aqueous medium.—II. Formation of double refractive " membranes " on gelatin globules by tannin.

† In chronological order.

DE JONG, H. G. BUNGENBERG, 1938 c. (**41**, 776.) Complex systems of biocolloids.—I. Discussion from the colloidchemical and from the electrochemical point of view.

DE JONG, H. G. BUNGENBERG, 1938 d. (**41**, 788.) Complex systems of biocolloids.—II. Specific factors of importance to the intensity of complex relations : tricomplex systems.

H. G. BUNGENBERG DE JONG with :

BANK, O., 1939 a. (**42**, 83.) Microscopic studies of biocolloids suspended in aqueous medium.—III. Coacervation phenomena in droplets of biocolloid sols enclosed in a collodion film : accumulation of dyes.

BANK, O., 1939 b. (**42**, 274.) Microscopic studies of biocolloids suspended in aqueous medium.—IV. Vacuolation phenomena in complex coacervate drops.

BANK, O., 1939 c. (**42**, 285.) Microscopic studies of biocolloids suspended in aqueous medium.—V. Gelatinized hollow spheres.

Miscellaneous papers

DE JONG, H. G. BUNGENBERG, 1936. *La coacervation et les coacervats* (*Actualités scientifiques et industrielles*, No. 397/398. Paris : Hermann et Cie). Monograph.

DE JONG, H. G. BUNGENBERG, 1937 a. *Kolloidzschr.*, **79**, 223 and 334. Review and discussion.

DE JONG, H. G. BUNGENBERG, 1937 b. *Kolloidzschr.*, **80**, 221 and 260. Review and discussion.

DE JONG, H. G. BUNGENBERG, BOOIJ, H. L. and WAKKIE, J. G., 1936. *Kolloid-Beihefte*, **44**, 254. Antagonism of neutral salts with regard to the reversal of charge of phospholipids.

DE JONG, H. G. BUNGENBERG and DEKKER, W. A. L., 1935. *Kolloid-Beihefte*, **43**, 143 and 213. Worked-out example of complex coacervation in system gum arabic and gelatin : viscosity analyses, influence of *p*H, salt resistance, etc.

DE JONG, H. G. BUNGENBERG and JOUKOVSKY, N. I., 1936. *C.R. soc. Biol.*, **123**, 149, 154, 299, 303 and 511. Series on reversal of charge of lecithin autocomplex coacervates by calcium chloride : film formation of lecithin, influence of cholesterol.

DE JONG, H. G. BUNGENBERG and TEUNISSEN, P. H., 1935. *Rec. Trav. chim. Pays-Bas*, **54**, 460. Microscopical measurement of electrophoretic velocity and determination of density of charge of hydrophilic colloids.

DE JONG, H. G. BUNGENBERG and TEUNISSEN, P. H., 1938. *Kolloid-Beihefte*, **47**, 254. Reciprocal hexol number and electrochemical equivalent weight of biocolloids.

DE JONG, H. G. BUNGENBERG, VERBERG, G. and WESTERKAMP, R. F., 1935. *Kolloidzschr.*, **71**, 184 and 194. Preparation and purification of clear lecithin sols : stability of lecithin sols.

DE JONG, H. G. BUNGENBERG and WINKLER, K. C., 1937. *Z. anorg. Chem.*, **232**, 119. Coacervation in inorganic salt solutions, analogous to complex coacervation in sols of hydrophilic colloids.

KELLEY, E. G., 1939. *J. Biol. Chem.*, **127**, 55 and 73. Interaction between dyes and cell substances.

KOETS, P., 1936. *J. Phys. Chem.*, **40**, 1191. Complex coacervation of amylophosphoric acid and proteins.

KOETS, P. and SCHOOFS, J., 1936. *Vlaamsch Natuurw. Tijdschr.*, **18**, 103. Coacervation used to determine effect of denaturation of ovalbumen.

KRUYT, H. R., 1933. *J. Int. Soc. Leath. Chem.*, **17**, 612. Colloid chemistry of collagen, gelatin and glue.

KRUYT, H. R., 1934. *Bull. Soc. chim. Belg.*, **43**, 559. Review.

KRUYT, H. R., 1937. *Bull. Soc. chim. Fr.*, **1925**. Part review.

KRUYT, H. R., 1939. *Chem. Weekbl.*, **36**, 375. Colloid chemistry of collagen, gelatin and glue.

LANGMUIR, I., 1938. *J. Chem. Phys.*, **6**, 873. The rôle of attractive and repulsive forces in the formation of tactoids, thixotropic gels, protein crystals and coacervates.

MODDERMAN, R. S. TJADEN, HOLLEMAN, L. W. J. and DE JONG, H. G. BUNGENBERG, 1934. *Kolloid-Beihefte*, **39**, 334. Simple coacervation of gelatin sols.

RUYSSEN, R. and VERSTRAETE, E. O. K., 1937. *Bull. Acad. Belg. Cl. Sci.*, **23**, 739. Coacervation in system gelatin—saponin.

PHYSICS AND TEXTILES

By A. B. D. CASSIE,
Wool Industries Research Association

§1. INTRODUCTION

PHYSICS and textiles in conjunction usually bring to mind X rays and long-chain molecules, for they are the subjects uppermost in the thoughts and publications of academic physicists. But the present *Report* records the more technical uses of physics in textiles. The problems of textiles are varied and often give interesting applications of well-established physical principles, one example being the interaction of our clothing with the heat and moisture contents of our changing atmosphere. On the other hand, many technical problems await the advance of fundamental physical theory; in particular the lack of any adequate theory of liquids and of co-operative phenomena in general puts a definite limit on technical advances in the realm of wet processing of textiles, and these processes are of the utmost importance to the industry. The *Report* deals with both solved and unsolved problems and indicates fundamental physical advances that would be welcomed in the industry.

The title of the *Report* is *Physics and Textiles*, and many of the subjects discussed are common to all textile fibres, but the reader will find a definite bias towards wool. The bias is doubtless largely due to the author's interest in wool, but in many respects wool is the most suitable of textile fibres for scientific study: its physical and chemical properties are almost independent of its geographical origin and, so far as has been determined, they are also reproduced in all mammal hairs and in horn, which provides a convenient form of keratin in bulk for studies that cannot be made with fibres. Cellulose, on the other hand, is a variable product with fewer well-defined properties, and the substance of the cotton fibre is not available in bulk. The materials of artificial fibres can be made in bulk form, but these fibres are now almost legion in number, and a review of their properties would require a book rather than a report.

§2. PHYSICAL CHARACTERISTICS OF CLOTHING TEXTILES

Clothing fabrics are usually woven from yarns which, in turn, are spun from fibres. They form approximately two-dimensional networks of yarns and ultimately of fibres. They owe their flexibility to their two-dimensional structure together with the fineness or small diameter of the constituent fibres. A wool felt composed of fibres oriented in all three dimensions may be no thicker nor heavier than a greatcoat fabric, but the three-dimensional network makes its resistance to bending too great for it to be used as clothing.

Clothing textiles, both natural and artificial, are hygroscopic, and a problem of interest to physicists, and of importance for textiles, is the interaction of the hygroscopic fibres with the water vapour of the atmosphere. The fineness of the fibres and their large total surface are significant features for the interaction

of a fabric with the atmosphere, and it is perhaps as well to outline these characteristics. Silk fibres vary from 2 to 6 μ in diameter; cotton is a ribbon rather than a cylinder, with an effective diameter of 10 to 20 μ; wool ranges in diameter from just under 20 to 40 μ (Matthews, 1923). Artificial fibres can, of course, be made of almost any diameter, but for clothing purposes they lie within the range covered by the natural fibres. The total surface in a given volume, V, of cylindrical fibres is given by $2V/r$, where r is the radius of the fibre, and for clothing fibres ranges from 2×10^4 sq. cm. per cc. for silk to 10^5 sq. cm. per cc. for the coarsest wool. A man's worsted suiting of 10 oz. per sq. yard* contains 215 cc. of wool fibre, and has a total fibre surface area of around 50 to 60 sq. yards per sq. yard of fabric.

The large surface area of the fibres would, however, be useless for interaction with the atmosphere if the fibres were closely packed together, and a second point for investigation is the degree of packing of the fibres. This is not uniform throughout the fabric, being greater in the yarns than in the interstices between them, and there is usually a fair amount of surface fibre which has a very low packing density. When allowance has been made for the surface fibre, typical values of the bulk density are 0·5 gm./cc. for fine worsted yarns, which are the most tightly packed yarns used in the wool industry, 0·43 for woollen yarns, whilst woven wool fabrics have densities ranging from 0·3 to 0·4 gm./cc. (Baxter and Cassie, 1945). As the density of the wool fibre is 1·3 gm./cc. (King, 1926), fine worsted yarns are 40% wool and 60% air. Wool has, however, a natural crimp that retains bulkiness in the yarn, and it is possible to make yarns of much greater bulk density from cotton. Values as high as 1·25 gm./cc. are not uncommon for cotton yarns, and as the fibre has a density of 1·5 gm./cc., the yarns contain up to 85% of cotton. Woven cotton fabrics have a density ranging from 0·25 to 0·9 gm./cc. (Peirce, 1930). Summing up, it may be said that the fibres in wool clothing have much access to the atmosphere, and that those in cotton have reasonable access to the air except in the case of high-density cotton yarns, where access to the air is doubtful.

§3. THERMAL PROPERTIES OF TEXTILES

The large percentage of air in textile fabrics has a profound effect on their thermal conductivities; in fact, the conductivity of fabrics is not very different from that for still air, and it does not depend greatly on the fibres forming the fabric. As the materials of natural fibres are not generally available except in fibre form, the values of their conductivities are usually deduced from their small contribution to fabric conductivity, and their values are not accurately known. Wool and horn are both formed from keratin, and it is likely that the conductivity of wool fibres is the same as that of horn, which Baxter (1946) gives, for heat flow normal to the direction of the long-chain molecules, as $4{\cdot}62 \times 10^{-4}$ cal./cm^2/sec. for a temperature gradient of 1° c. per cm. at 1% regain,† and $5{\cdot}32 \times 10^{-4}$ at 11% regain; the relationship between conductivity and regain is roughly linear. The cellulose material of cotton is not available in bulk form. Regenerated cellulose is, of course, available, but its physical properties do not appear to be at

* Cloth is usually sold per yard, 60 inches wide, and weighs 16 to 20 oz. per yard.

† Regain is the moisture content expressed as a percentage of the dry weight.

all similar to those of cotton fibres ; in particular, the water absorption of regenerated cellulose is very different from that of cotton. The material of silk, fibroin, is also not available in bulk form.

Schuhmeister (1877) and Rubner (1895) used a formula for the conductivity of heterogeneous materials to estimate the conductivities of fibres ; the formula appears to have been given to them by Stefan without publication, but it fits experimental data for textiles reasonably well. It is

$$k_m = \tfrac{1}{3}(k_1 v_1 + k_2 v_2) + \tfrac{2}{3}\frac{k_1 k_2}{k_1 v_2 + k_2 v_1}, \qquad \ldots\ldots(1)$$

where k_m is the conductivity of the mixture, k_1 is the conductivity of the component of fractional volume v_1, and k_2 is that of the component of fractional volume v_2. It is easily seen that equation (1) is derived by assuming that one-third of the fibres form slabs parallel to the direction of heat flow and two-thirds form slabs normal to this direction. Schuhmeister and Rubner both used fibres randomly

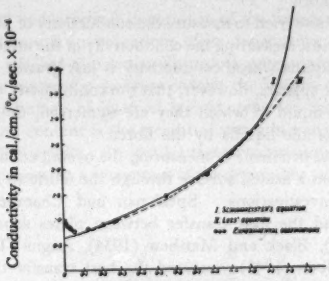


Figure 1. Conductivity of wool felts.

oriented in measuring k_m, and Rubner gives the conductivities of the fibres as $4{\cdot}8 \times 10^{-4}$ cal./cm./°c./sec. for mammal hairs, $8{\cdot}9 \times 10^{-4}$ for silk, and $14{\cdot}2 \times 10^{-4}$ for plant fibres.

Baxter (1946) has made a careful study of the thermal conductivity of wool felts at bulk densities ranging from 0 to 1 gm./cc. This range of densities far exceeds that of Schuhmeister and Rubner, who did not extend their measurements beyond 0·1 gm./cc. Baxter used Hercus and Sutherland's (1934) method of measuring thermal conductivity ; his results are shown in figure 1. Equation (1) fits his data excellently and extrapolates to the value obtained for horn when the fractions of fibres lying parallel and normal to the direction of heat flow are taken as 0·29 and 0·71 respectively.

Baxter (1946) has also fitted his data to the equation,

$$k_m = k_1^{v_1} k_2^{v_2}, \qquad \ldots\ldots(2)$$

proposed by Lees (1900) for the flow of heat normal to the length of long prisms of square cross-section, the prisms of conductivity k_1 and k_2 and fractional volumes

v_1 and v_2 being arranged in a draught-board pattern. There is little to choose between equations (1) and (2) so far as the experimental data for felts are concerned, but equation (1) with the coefficients 0·29 and 0·71 gives more reasonable values for the fibre conductivities. Baxter concludes that Rubner's value for the conductivity of wool is roughly correct, but that the value for cotton may be as much as 30% too large.

A peculiar feature of Baxter's data is the increase of conductivity at very low bulk densities. It is probably due to convection currents set up in the large air cells of the low-density felts by the more rapid conduction of heat along the fibres than through the air ; Forbes (1872) apparently noticed a similar phenomenon, for he states that the conducting power of cotton decreases with compression. The existence of a minimum in the conductivity-density curve near 0·1 gm./cc. makes Schuhmeister and Rubner's results somewhat doubtful, for they worked exclusively in this region, and it is surprising that Rubner should give a value for the conductivity of wool, $4{\cdot}8 \times 10^{-4}$ cal./cm./°c./sec., which lies so close to that for horn.

Baxter (1946) has also tried to measure the conductivity of fibres by immersing them in different liquids, measuring the conductivity of the mixtures, and bracketing the fibres with liquids whose conductivity is just greater and just less than that of the fibres. It appears, however, that the conductivity of the fibres varies very much with the liquid in which they are immersed, doubtless due to the variable absorption of the liquids by the fibres.

The more practical problems of measuring the overall conductivity of fabrics and the heat flow from a heated surface through the fabric to air have been the subjects of many investigations. Speakman and Chamberlain (1930) and Rees (1941) measured the heat transfer between plates separated by fabrics, whilst Marsh (1931), Black and Matthew (1934), Angus (1936), Niven and Babbitt (1938) and Rees (1941) measured the heat transfer through fabrics to air. The heat transfer from a hot surface through the fabric to air, is, of course, the quantity required for clothing and many heat-insulation problems. It is not readily subject to accurate measurement because it is difficult to define the thickness of a fabric : a few surface fibres that would not resist deformation under the lightest loads can maintain a still-air layer at the exposed surface, and so materially affect the heat loss ; and equally, the thermal resistance between the hot surface and the fabric is very sensitive to the load; or weight of the fabric. Baxter and Cassie (1943) proposed that the thickness of a fabric should be arbitrarily taken as its thickness under a load of 5 gm./cm^2, and that any effective thickness in excess of this measured value should be included in the surface emissivity of the fabric. Using this convention, and measuring the heat loss through a number of layers of fabric, they evaluated the thermal conductivities and emissivities of fabrics. The thermal conductivities of clothing fabrics average around 1×10^{-4} cal./cm./°c./sec., and the surface emissivities in ambient air around $2{\cdot}5 \times 10^{-4}$ cal./cm^2/°c./sec. ; the surface resistance thus effectively increases the thickness of the fabric by 0·4 cm., and as this is much greater than the thickness of many fabrics, it is seen that the surface resistance and the air layers between fabrics are much more important than the fabrics themselves for heat insulation in ambient air. In moving air, the surface emissivities

increase. Baxter and Cassie (1943) give the emissivity of blue serge as $2{\cdot}6 \times 10^{-4}$ cal./cm²/°c./sec. in ambient air and as $7{\cdot}5 \times 10^{-4}$ in air moving at 4 m.p.h., parallel to the cloth surface, whilst a specimen of blue blanket gave values of $1{\cdot}9 \times 10^{-4}$ and $3{\cdot}4 \times 10^{-4}$ under the two different conditions. Thus, even under moving air conditions, the surface resistance of the blanket is equivalent to a blanket thickness of almost 0·3 cm. It should be noted from these data that the surface fibres of the blanket are effective in maintaining a low emissivity under conditions of moderate air movement.

Dietz (1912) has made an extensive study of the specific heat of textiles. The values for dry animal and vegetable fibres lie within the range 0·32 to 0·33, and fibres containing moisture have values given by the additive formula. Dietz states that the influence of temperature is negligible between 0° and 100° c.

§4. EQUILIBRIUM MOISTURE RELATIONS FOR TEXTILES

The commercial and practical importance of the moisture content of textiles has led to many thorough investigations of its dependence on temperature and relative humidity. It has also led to an unfortunate confusion of teminology. The textile worker means by *moisture content* the weight of water in a textile expressed as a percentage of its total weight, whereas the accepted scientific meaning of moisture content is the weight of water in a material expressed as a percentage of its dry weight. In textiles the term *regain* is used to denote the weight of water expressed as a percentage of the dry weight, and to avoid confusion

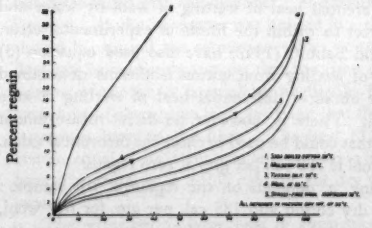


Figure 2. Regain : relative humidity curves.

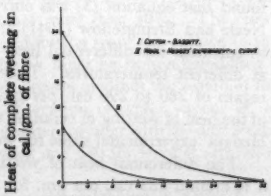


Figure 3. Heat of complete wetting.

it will be used throughout this *Report*. The greatestuncer tainty in regain determinations is due to the difficulty of obtaining a constant dry weight ; it is usually $\frac{1}{2}$ to 1% greater when dried at 25° c. in a vacuum or in dry air than it is when dried at 100° c. The B.S.I. Standard (1942) specifies air at 102 to 105° c. for determination of the dry weight of textiles.

Typical regain-relative humidity curves are shown in figure 2. They are all hysteresis loops, as shown for wool, and the other curves reproduced are for absorption : the regain on desorption is greater than that on absorption. Schloesing (1893) appears to have made the first determinations for silk, wool,

and cotton, and more recent data, which show the hysteresis loops, are given by Denham and Allen (1933) for silk, by Speakman (1936) for wool, and by Urquhart and Williams (1924) for cotton. In general, the water absorption of animal hair, horn, and other forms of keratin is closely similar to that of wool, and natural cellulose products usually show regains similar to those of cotton. The artificial cellulose fibres have variable regain curves depending on their composition and method of production.

The thermodynamics of water absorption was first discussed by Trouton and Pool (1906), who concluded that if the heat of absorption of water by the fibres was equal to the latent heat of evaporation of water in bulk, then the regain-relative-humidity curve should be independent of temperature. Shorter (1924) showed that the absorption of liquid water by textiles was accompanied by considerable evolution of heat, and applied Kirchhoff's law relating the heat of dilution of a solution to its vapour pressure to the problem of absorption of water by textiles. Kirchhoff's relation is

$$\frac{\partial}{\partial T}\left(\log \frac{p}{p_0}\right) = \frac{q}{RT^2}, \qquad \ldots\ldots(3)$$

where p is the vapour pressure in equilibrium with the textile, p_0 is the saturation water vapour pressure, and q is the differential heat of wetting, or is the heat evolved per mole of liquid water added to an infinite mass of textile. As $100\,p/p_0$ is the relative humidity of the air, equation (3) gives the change of the equilibrium relative humidity with temperature at a constant regain.

Shorter used equation (3) to deduce differential heats of wetting for wool and cotton, using Schloesing's data on regain, relative humidity, and temperature. Hedges (1926) measured the differential heat of wetting of wool by water and found that equation (3) was correct to within the limits of experimental error. Neale and Stringfellow (1941) and Babbitt (1942) have also used equation (3) to calculate the differential heat of wetting from cotton isotherms determined at different temperatures. They obtain a differential heat of wetting at zero regain of 280 to 290 cal. per gm. There is, however, no direct measurement of the heat of wetting of cellulose that could be used to check the theoretical value. Hedges' experimental value for wool is 185 cal. per gm. at zero regain.

The differential heat of wetting, q, depends on the regain of the sample: it is around 285 cal. per gm. for dry cotton and 185 cal. per gm. for dry wool, and it becomes zero for textiles exposed to saturated air. Figure 3 shows the total heat of wetting, or the heat evolved when 1 gm. of fibre at a given regain is immersed in liquid water The curve for wool is Hedges' (1926) experimental data, and that for cotton is from Babbitt's (1942) calculated differential heats of wetting. The cotton values may err considerably, but they show that, despite the large heat of sorption for dry cotton, the total heat of wetting is small compared with that for wool.

§5. TEXTILES AND HUMID ATMOSPHERE CHANGES

The rate of change of regain when the atmosphere around a textile is changed is almost as important as its equilibrium value. Experience has long since established that a change of atmosphere is only very slowly followed by textile materials, and it has generally been assumed that the slow change was due, in

part at least, to a slow diffusion of water vapour from the surface to the interior of the fibres. If, however, one estimates the time required for this diffusion from the dimensions of the fibre and likely values of the diffusion coefficient, it becomes apparent that diffusion into the fibre cannot be the rate-controlling process. Hill (1928) gives convenient numerical solutions of the diffusion equation for diffusion into an infinite cylinder when the surface concentration is suddenly changed. The time, t, required for the average concentration within the fibre to reach 80%, where 100% represents the total change in concentration, is given by the relation

$$\frac{kt}{r^2} = 0{\cdot}2,$$

where k is the diffusion coefficient and r is the radius of the cylinder. The average radius ofa tex tile fibre is rather less than $10\,\mu$, or 10^{-3} cm., so that kt is of the order of 2×10^{-7}, and t will be no more than 2 sec. for a diffusion coefficient of 10^{-7} cm.^{2}/sec. As the time required for change of regain with change of atmosphere conditions ranges from hours to weeks, depending on the bulk of the material, the time of diffusion of water vapour from the surface to the interior of the fibre must be of little significance.

Trouton and Pool (1906) remark that " textile fibres absorb water vapour vigorously ", and presumably they mean that it is absorbed rapidly by the fibres, but later work (Fisher, 1923) suggests the contrary. King and Cassie (1940) have recently reinvestigated the problem of the rate of absorption of water vapour by wool fibres. A mass of roughly 0·25 gm. of wool was suspended by a sensitive spring balance in an evacuated chamber, and water vapour at 23·5 mm. pressure was introduced into the chamber ; the resulting regain-time curve is shown in figure 4. As the apparatus was housed in a thermostat at 25° c., the final regain should be around the saturation value, or rather more than 30%. The curve does not immediately suggest that the time of diffusion from the surface to the interior of the fibres is small or negligible, but it must be remembered that a large heat of absorption is evolved and that this greatly modifies the regain of the outermost layer of the fibre. The heat of absorption of water vapour is, of course, equal to the latent heat of evaporation of water together with the differential heat of wetting, or it is 765 cal./gm. for dry wool and 580 for wool at saturation regain.

The effect of the large heat of absorption is clearly shown in figure 5. It was obtained by winding fine platinum wire into the wool and using this as a resistance thermometer. The large increase in temperature means, according to Kirchhoff's relation, that the regain must be small at 23·5 mm. pressure, even though the time for diffusion into the fibre is negligible. King and Cassie (1940) assumed that the time for diffusion was, in fact, negligible, and used Kirchhoff's relation to determine the temperature of the wool on first exposure to the water vapour ; the calculated initial temperature was 80° c., and the maximum regain that could be instantaneously attained was shown to be 2·4%. They assumed that the rate-controlling process after introduction of the water vapour was cooling and, using Newton's law, they calculated the increase of regain and decrease of temperature with time. The calculated and observed values are compared in figures 4 and 5, and it is seen that the assumption of a

negligible time for diffusion into the fibres is fully justified. The one adjustable parameter in the calculations is the emissivity of the wool sample. It is difficult to assign an accurate surface area to the sample, but reasonable estimates give emissivities between 2 and 3×10^{-4} cal./cm^2/° c./sec., whilst Preston (1929) quotes values ranging from 2·5 to 3×10^{-4} for black surfaces, with temperature differences up to 30 °c. between the surface and its surroundings. The adjustable parameter in the calculation thus has a reasonable value.

The result that textile fibres are for practical purposes always in equilibrium with the atmosphere immediately surrounding them has a number of interesting consequences for clothing problems. The fibres in a fabric are accessible to the atmosphere, and it has been shown that the heat insulation of fabrics is largely due to the air entrapped between them. The propagation of a temperature change will therefore be largely determined by the diffusivity of the air entrapped in the fabric, and it happens that around normal temperatures the diffusivity of

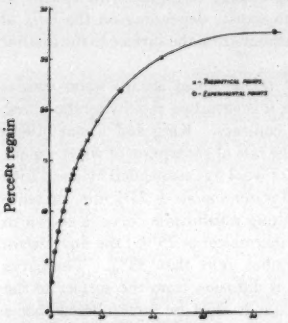


Figure 4. Rate of absorption of water vapour by wool.

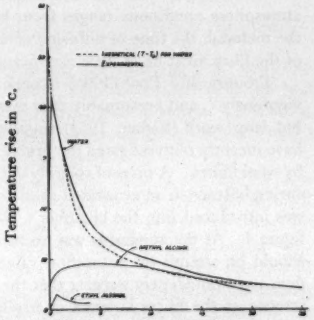


Figure 5. Temperature change on absorption of vapours by wool.

(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

air, 0·21 cm^2/sec., is close to the diffusion constant of water in air, 0·24 cm^2/sec. The phenomena of temperature propagation and diffusion of water vapour into a fabric must therefore follow one another closely, and the large heat of sorption of water vapour by the fibres must have a profound effect on temperature changes that occur in the fabric.

The large amount of water sorbed by textiles and the large heat of sorption of water combine to make the overall heat change associated with changes in atmosphere conditions of real importance (Cassie, Atkins and King, 1939). Thus, the change that occurs on proceeding from a well-ventilated room indoors to outdoors is largely a decrease of temperature : the water vapour pressure will remain roughly constant. A typical example is the change from indoor conditions of 18° c. and 45% R.H. to outdoors at 5° c. and 95% R.H. The equilibrium regain of wool clothing under the first set of conditions is 10%, and under

the second is 27%. Every kilogram of wool brought from indoors to outdoors must therefore absorb 150 gm. of water, and as the heat of sorption is around 650 cal./gm., the total heat evolved during absorption of the water is 97,500 cal. A man's suit weighs roughly 1½ kilograms, and the heat evolved by it will be roughly 150,000 cal., or as much as his body normally loses in 1½ hours. By contrast, the specific heat of the suit contributes 6000 cal. towards protection against the change of temperature. The interaction of hygroscopic textiles with the moisture of the atmosphere thus gives much protection against sudden changes of temperature ; the textiles have, in effect, an enormous specific heat. Equally, carpets and textile furnishings in a room will tend to maintain a constant temperature, and this may give rise to their reputation of "warmth".

A quantitative investigation of the interaction of atmospheric humidity and textiles is probably best approached by extension of the theory of the wet-bulb hygrometer (Cassie, 1940a). The fibres may be regarded as a large source of water, just as the wick of a wet-bulb hygrometer is assumed to be an infinite source of water, and August's hypothesis (*cf.* Whipple, 1933) that the air leaves the wet-bulb saturated at the wet-bulb temperature, and that the heat required to evaporate the water comes from the same sample of air as that into which the water is evaporated, can then be applied to the textile problem. The direct application of August's hypothesis ignores the heat capacity of the textile, but it has little effect compared with the large heat of sorption of water.

Suppose, then, that a mass of textile fibres is in equilibrium with air of temtemperature T_0 and water vapour concentration C_0, and that air of temperature T_1 and water vapour concentration C_1 is forced through it. The issuing air will have some temperature T_w and water vapour concentration C_w determined by the heat balance of August's hypothesis, or by

$$\rho_a c_a(T_1 - T_w) = \gamma(C_w - C_1), \qquad \ldots\ldots(4)$$

where ρ_a is the density of the air ; c_a is its specific heat, and γ is the heat of sorption per gm. of water ; its average value may be taken as 650 cal.

Equation (4) is one relation between T_w and C_w ; another is obtained from the assumption that the regain of the textile does not change appreciably whilst the wet-bulb temperature is established. At constant regain the temperature and water-vapour concentration in equilibrium with the textile must satisfy Kirchhoff's relation, or it must satisfy

$$\frac{1}{C}\frac{\partial C}{\partial T} = \frac{\Gamma}{RT^2} - \frac{1}{T}, \qquad \ldots\ldots(5)$$

where Γ is the heat of sorption per mole of water and R is the gas constant per mole.

Take as an example a change of temperature. A textile in equilibrium with air of 20° c. and 60% R.H., or water vapour concentration $10{\cdot}5 \times 10^{-6}$ gm./cc., has air at 35° c. and water vapour concentration $10{\cdot}5 \times 10^{-6}$ gm./cc. forced through it. The temperature change, ΔT is 15° c., and (1) may be written

$$\rho_a c_a(\Delta T - \delta T) = \gamma \delta C, \qquad \ldots\ldots(6)$$

where δT is $(T_w - T_0)$ and δC is $(C_w - C_0)$. Equations (5) and (6) now determine δT and δC in terms of the given temperature change ΔT. $\rho_a c_a$ is $2{\cdot}9 \times 10^{-4}$ cal./cm.^{3}/°c. at 25°c. and 760 mm. pressure; δT is found to be 6° and

δC is 4×10^{-6} gm./cm.^{3} The temperature of the air issuing from the textile is therefore 26° c. and its water vapour concentration is $14{\cdot}5\times10^{-6}$ gm./cm.^{3}

The hygroscopic properties of the textile thus decrease the change in temperature from 15° c. to 6° c.; when sufficient air has passed through it, the regain will decrease, the temperature of the issuing air will increase, and will ultimately approach the value, 35° c., for the incident air.

A second interesting example is that of an increase of water vapour concentration. If the textile is again conditioned to 20° c. and 60% R.H. and the humidity is increased to 90%, or to $15{\cdot}8\times10^{-6}$ gm./cc., equation (4) may be written

$$\rho_a c_a \delta T = \gamma(\Delta C - \delta C), \qquad \ldots\ldots(7)$$

where ΔC is $(C_1 - C_0)$ or $5{\cdot}3\times10^{-6}$ gm./cm.^{3} Equations (5) and (7) now give δT as 4·8° c. and δC as $3{\cdot}1\times10^{-6}$ gm./cm.^{3} The temperature of the textile and of the issuing air thus increase to 24·8° c., and the water vapour concentration decreases from $15{\cdot}8\times10^{-6}$ to $12{\cdot}7\times10^{-6}$ gm./cm.^{3} It is an example of the wet

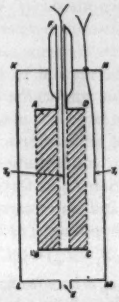


Figure 6. Measurement of propagation of temperature change through textiles.
(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

bulb in reverse; water vapour is absorbed from the stream of air, and the heat of absorption which is evolved raises the temperature of the air to give a negative "wet-bulb depression".

Baxter and Cassie (1940*, 41) have designed an experiment to check this theory of propagation of temperature change through textiles. They used a hollow cylinder, ABCD of figure 6, made entirely of textiles, through which air could be forced mechanically. The outer and inner walls of the cylinder were made of woven starched linen, and between these walls fibres were packed to a density of 0·15 gm./cm.^{3} This cylinder was housed in an outer brass cylinder, KLMN, which could be transferred rapidly from a thermostat at 20° c. to one at 35° c. Air was forced through coils in the thermostats and then through the opening, E, into the outer cylinder, after which it filtered through the mass of textile to escape through the hollow centre of the inner cylinder. Thermocouples, T_2 and T_1, measured the temperature of the air before and after passing through the mass of textile fibres. The water vapour concentration was maintained at $10{\cdot}5\times10^{-6}$ gm./cm.^{3} throughout the experiments.

* Cassie and Baxter, 1940 in references.

Temperature-time curves are shown in figure 7 for representative natural fibres. The curves show the temperature of the air before and after passing through the textile from the time of transfer of the apparatus from the thermostat at 20° to that at 35° C. The curves all show a rapid increase of temperature from 20 to 26° C. This is in excellent agreement with the temperature determined from the wet-bulb theory for the conditions of the experiment, which were those used in the example of the preceding paragraphs. When the wet-bulb temperature has been attained, the temperature of the exit air gradually approaches that of the incident air, the fibres whose regains change more rapidly with relative humidity showing the slower approach to the final temperature.

Theories have been given which describe the full temperature-time curves. Henry (1939) gave the theory for diffusion of humid air into hygroscopic masses,

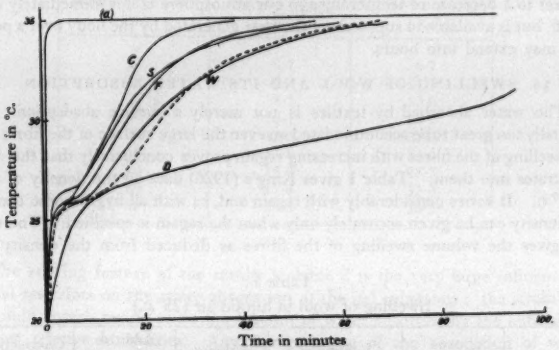


Figure 7. Rate of propagation of temperature change through textiles.
(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

Curve (*a*). Air temperature before passing through textile cylinder.
Curves *C*, *S*, *F*, *W*. Air temperature after passing through cylinders of cotton, silk, flax, and wool respectively.
Curve *D*. Diffusion through cotton cylinder.
------ Theoretical curve for wool.

and it was checked with the apparatus that has just been described. The central exit tube, F of figure 6, was closed and the air from E was allowed to escape from the top of the outer cylinder; temperature changes had then to diffuse through the textile to the central thermocouple T_1. The temperature-time curves for diffusion also show the discontinuity at the wet-bulb temperature, and the change to the final temperature is, of course, much slower.

Cassie (1940 b) gave the theory for air forced through a mass of textile ignoring all diffusion effects. Under these circumstances, the temperature change is propagated in the form of two wave fronts: the first sets up the wet-bulb conditions and the second the final conditions. The first wave front travels at roughly 30 times the speed of the second front through wool packed to a density of 0·1 gm./cm.^{3} Daniels (1941) took into account diffusion of heat and water

vapour in the air stream forced through the textile, and the broken curve of figure 7 shows the theoretical temperature-time curve for wool at a bulk density of 0·1 5gm./cm^3 under the conditions of the experiment ; the air speed through the textile was 1·75 cm./sec. The excellent agreement of theory with experiment probably forms the best proof of the assumption that, for all practical purposes, the fibres are always in equilibrium with the atmosphere surrounding them.

A point of practical importance is the time required for the change of regain. Figure 7 shows that when a temperature change diffuses into the mass of fibres, the time required for significant changes of regain is of the order of 1–2 hours. This means that the large amount of heat evolved by our clothing on proceeding from indoors to outdoors is effective over a period of hours rather than minutes. Thus, the large amount of heat generated by clothing when subject to a decrease of temperature in our atmosphere is not immediately dissipated, but is available to supplement the heat generated by the body over a period that may extend into hours.

§6. SWELLING OF WOOL AND ITS WATER ABSORPTION

The water absorbed by textiles is not merely a surface absorption : it is generally too great to be accommodated on even the large surface of the fibres, and the swelling of the fibres with increasing regain proves conclusively that the water penetrates into them. Table 1 gives King's (1926) data on the density of wool at 25° c. It varies considerably with regain and, as with all hygroscopic textiles, the density can be given accurately only when the regain is specified. The table also gives the volume swelling of the fibres as deduced from the density and

Table 1

Swelling of wool in humid air (25° c.)

Regain, r (%)	Density (gm./cm^3)	Volume of $(100+r)$ gm. of wool (cm^3)	Volume swelling (%)	Additive volume of $(100+r)$ gm. (cm^3)	Contraction in volume (%)
0·0	1·304	76·8	—	76·8	—
2·0	1·311	77·8	1·3	78·5	0·9
5·0	1·318	79·8	4·0	81·5	2·1
7·0	1·318	81·2	5·8	83·5	2·8
10·0	1·317	83·5	8·8	86·5	3·6
15·0	1·316	87·4	14·0	91·5	4·7
20·0	1·308	91·7	20·0	96·5	5·3
25·0	1·300	96·2	26·0	101·5	5·5
30·0	1·281	101·4	32·0	106·5	5·0
33·0	1·270	104·7	36·0	109·5	4·6

weight of $(100+r)$ gm. of wool. If the volume swelling is assumed to be wholly radial, the swelling given in table 1 agrees to within the limits of experimental error with the radial swelling obtained from microscope measurements (Barker, 1929). The measured longitudinal swelling is, in fact, negligible. The last two columns of table 1 show the volume of 100 gm. of wool and r gm. of water, and the contraction in volume when the water is absorbed by the wool.

The limited swelling of textile fibres in water means that the mechanical strength of the fibres prevents complete solution. There is, in fact, a natural restraint on swelling, and this subjects the absorbed water to compression, which in turn considerably modifies the water-vapour-absorption isotherm. Barkas (1942 a, 42 b) has discussed fully the influence of natural restraints on the swelling pressure and absorption isotherms of wood. He has calculated the regain of the gel substance of spruce wood under its natural restraints and in a stress-free condition. His results are given in table 2.

Table 2

Regain of gel substance of spruce wood

Relative humidity (%)	Natural regain (%)	Stress-free regain (%)
10	3·7	6·7
20	6·3	11·2
30	8·3	15·0
40	10·1	18·1
50	12·0	21·0
60	13·8	24·6
70	15·7	28·3
80	17·8	32·0
90	20·3	36·5

The striking feature of the results in table 2 is the very large influence of natural restraints on the water absorption of the gel substance : the stress-free gel would absorb nearly twice the amount of water absorbed by the natural gel at most relative humidities. Any intrepretation of the absorption of water by gels must therefore be based on stress-free isotherms rather than on natural isotherms.

Wood is a complex aggregate of long cylindrical cells of gel substance with a hollow centre and encased in restraining sleeves, and Barkas's analysis has been developed to deal with this aggregate. Textile fibres, and wool in particular, are much more homogeneous, and it is possible to use a simpler analysis in determining the stress-free isotherm. Barkas (1942 a) has shown that only hydrostatic pressures can affect the vapour pressure in equilibrium with a gel, and that shear stresses do not influence the isotherm. It is therefore necessary to determine only the hydrostatic pressure developed by the natural restraints.

The wool fibre is symmetrical about the fibre axis, and its elastic constants will be isotropic in the plane normal to its axis. The fibre does not swell longitudinally when absorbing water, and as the hydrostatic pressure, P, produces a longitudinal deformation $P(1/E_l - 2\sigma_{rl}/E_r)$, where E_l and E_r are Young's moduli for longitudinal and radial deformations, and σ_{rl} is Poisson's longitudinal ratio for radial loading, this expression must be zero, or

$$\frac{1}{E_l} - 2\frac{\sigma_{rl}}{E_r} = 0. \qquad \ldots\ldots(8)$$

The radial deformation, $\delta r/r$, under P is given by

$$\frac{\delta r}{r} = \frac{P}{E_r} - \sigma_{lr} - \sigma_{rr}\frac{P}{E_r}, \qquad \ldots\ldots(9)$$

where σ_{lr} is Poisson's radial ratio for longitudinal loading and σ_{rr} is the radial ratio for radial loading.

Speakman (1927) has given load-elongation curves for wool fibres under the full range of regains, but apart from his data little is known of the elastic constants of wool or keratin. Nothing is known of E_r, but it can be eliminated from equations (8) and (9) to give

$$\left(\frac{1}{2\sigma_{rl}} - \sigma_{lr} - \frac{\sigma_{rr}}{2\sigma_{rl}}\right)P = E_l\frac{\delta r}{r} \qquad \ldots\ldots(10)$$

and, knowing the radial swelling of the fibre, the right-hand side of equation (10) can be evaluated from Speakman's curves. There remains, however, the difficulty of the unknown Poisson's ratios, or the factor

$$\left(\frac{1}{2\sigma_{rl}} - \sigma_{lr} - \frac{\sigma_{rr}}{2\sigma_{rl}}\right).$$

An estimate of this factor can be made by using the fact, which will be discussed later in the *Report*, that at low regains the water is largely absorbed on to low-energy localized sites, and that the differential heat of wetting should be a constant at low regains. The measured differential heat of wetting, q, discussed in §4, is not constant at low regains, but it must be remembered that the measured values correspond with the heat of absorption on to the low-energy sites less the swelling energy. If the swelling energy, PV, where V is volume of unit mass of absorbed water, is added to the measured differential heats of wetting, these should remain constant at low regains. The unknown factor in equation (10) is therefore chosen to make $(q+PV)$ a constant at low regains. Following this procedure, the factor $(1/2\sigma_{rl} - \sigma_{lr} - \sigma_{rr}/2\sigma_{rl})$ is found to lie near 0·5, and

$$P = 2E_l\frac{\delta r}{r}. \qquad \ldots\ldots(11)$$

It is worth noting that if the fibre is treated as a completely isotropic material of Poisson's ratio 0·25, which is a usual value for solids, then the ordinary bulk modulus relation gives :

$$P = \frac{E}{3(1-2\sigma)}\frac{\delta V}{V} = \frac{E}{(1-2\sigma)}\frac{\delta r}{r}, \qquad \ldots\ldots(12)$$

where E is Young's modulus for the material and σ is Poisson's ratio; it becomes, when σ is 0·25,

$$P = 2E\frac{\delta r}{r} \qquad \ldots\ldots(13)$$

which is identical with equation (11) when E is identified with E_l.

At high regains there is no method of evaluating $(1/2\sigma_{rl} - \sigma_{lr} - \sigma_{rr}/2\sigma_{rl})$, and to make calculations possible, it is assumed that it retains the value 0·5.

When the hydrostatic pressure is known, the stress-free isotherm can be readily determined by use of the usual thermodynamic relation :

$$P = \frac{RT}{Mv}\log\frac{h_0}{h}, \qquad \ldots\ldots(14)$$

where h_0 is the observed vapour pressure and h is the vapour pressure that

would be given by stress-free fibres, Mv is the molar volume, R is the gas constant, and T is the absolute temperature.

Cassie (1945 b) has used relations (11) and (14) to obtain a stress-free isotherm for wool. The analysis is given in table 3, where it may be seen that the hydrostatic stress becomes very great, particularly around 10% regain. At regains greater than 10%, the load at a given elongation decreases so rapidly that the hydrostatic stress decreases despite the increased swelling. Figure 2 shows the normal and stress-free isotherms, and it may be seen that just as for the gel substance of wood, the stress-free regain of wool at a given vapour pressure is around twice that of the normal regain.

Barkas (1942 b) has suggested that the hysteresis shown by absorption and desorption isotherms is largely a reflection of the mechanical hysteresis of the gel. A straightforwardap plication of equation (14) to the absorption-desorption isotherms indicates that this hypothesis is probably correct for wool. At a given regain, or at a given swelling, the hydrostatic pressure on the absorbed water will be greater during absorption than during desorption, and the observed vapour

Table 3

Hydrostatic stress and hysteresis for 25° c. wool-water isotherm.

Regain (%)	P (gm./cm² × 10⁵)	PV (cal./gm.)	$(q+PV)$ cal./gm.	$\frac{h_0}{h}$	h_0 (mm. of mercury)	h (mm. of mercury)	H_1 %	H_2 %	δP (gm./cm² × 10⁵)
5	16	38	187	3·15	2·9	0·92	13	9	5·4
7	22	53	187	4·85	5·0	1·03	25	16	4·6
10	22	53	164	4·85	8·4	1·74	40	34	2·2
15	21	51	120	4·5	14·4	3·2	68	61	1·4
20	16	38	58	3·15	18·6	5·9	87	82	0·8
25	14	34	2	2·75	21·0	7·7	95	93	0·5
30	12	—	—	2·35	22·7	9·6	—	—	—
33	10	—	—	2·15	23·8	11·6	—	—	—

pressure at a given regain will be greater during absorption than during desorption. This qualitative description is in accord with Speakman's (1936) experimental observations. A semi-quantitative analysis may be made by use of equation (14). If, at a given regain, the relative humidity of the air in equilibrium with the absorption isotherm is H_1, and that in equilibrium with the desorption isotherm is H_2, equation (14) gives for the difference, δP, in hydrostaticstress on absorption and desorption

$$\delta P = \frac{RT}{Mv} \log \frac{H_1}{H_2}. \qquad \ldots\ldots(15)$$

The last column of table 3 gives the values of δP determined from the hysteresis curves for wool. They are seen to be of the same order of magnitude as the values of P, although these have been calculated only from known elastic constants of wool. The values of δP are less than those of P, and fit in very well with the hypothesis that sorption hysteresis is a reflection of mechanical hysteresis. δP increases rapidly as the regain decreases, which is in accord with the poor recovery of wool at low regains from deformations imposed at higher regains, a phenomenon known as "temporary set".

§7. THE PROBLEM OF MULTIMOLECULAR ABSORPTION

Multimolecular absorption, or the simultaneous absorption of one species of molecule in two or more forms, is a phenomenon of exceptional importance in the textile and, indeed, in many industries. It is a problem that has led to much controversy, and until the fundamental physical principles involved in multimolecular absorption are clearly understood, many technical problems must remain unsolved.

The isotherms for multimolecular absorption are usually sigmoid in shape. Katz (1917) recorded an extensive series of experimental isotherms for water ranging from absorption by crystals, cellulose, and keratin to solution by sulphuric acid and glycerine. They are all of sigmoid shape, and the absorption and solution were accompanied by large heat effects. This type of isotherm is not, however, restricted to water: inert gases, or any gas, may show sigmoid absorption isotherms at temperatures below their critical temperature (*cf.* Brunauer and Emmett, 1937).

Katz (1917) was most insistent that multimolecular absorption was a form of solution, although it should be noted that absorption by a solid lacks the essential requirement of a solution that the two component molecules should be able to interchange positions. This distinction between absorption and solution gave favour to the capillary pore theory (*cf.* Hedges, 1926) in which the gases were supposed to condense in capillaries at vapour pressures less than their saturation vapour pressures because of the concave surface of the liquid in the pore. The capillary pore hypothesis cannot, however, be called a theory of absorption, for it does not suggest how a sufficient number of molecules can condense in a dry pore to form the liquid surface required for depression of the vapour pressure. There is no doubt that the presence of capillaries will have a marked influence on the vapour pressure of condensed gases, but their presence cannot explain the initial condensation.

The present conception of multimolecular absorption is that localized sites, as in Langmuir's monomolecular absorption, exist for the initial stages of absorption; further absorption is assumed to be due to some form of condensation on the molecules already in Langmuir's layer. There is agreement that this is the correct general view, but the mechanism of condensation causes much trouble.

De Boer and Zwikker (1929) and Bradley (1936) considered that the condensation of multilayers was due to polarization and orientation of the dipoles of absorbed atoms or molecules by the absorbent. They deduced essentially the equation

$$\log p_n = \alpha\lambda^{2n} \qquad \ldots\ldots(16)$$

for the pressure p_n in equilibrium with the nth layer, α is a constant, and λ is the ratio of the dipole moment in the nth layer to that in the $(n-1)$th layer. Equation (16) can be made to fit the sigmoid curve of multimolecular absorption, and Bradley (1936) has used it to explain the isotherms of argon absorbed on copper and aluminium sulphate. The experimental value of λ is, however, very close to unity, being 0·99 for argon absorbed on aluminium sulphate. Brunauer, Emmett and Teller (1938), have, on the other hand, estimated the value of λ from the refractive index of argon, and find that 0·01 rather than 0·99 would be expected from the polarizability and dimensions of the argon atom.

It seems, therefore, that the hypothesis of dipole moments induced from layer to layer is unable to explain the condensation quantitatively ; even in the second layer the polarization energy for argon is only 10^{-4} of that in the first layer.

Brunauer, Emmett and Teller reject the polarization hypothesis because the energies involved are much too small. They postulate instead binding forces similar to those in the liquid state for all but the first layer, which they assume to be absorbed on to localized sites as in Langmuir's monolayer. They use this hypothesis together with an evaporation-condensation mechanism to derive an isotherm equation. Their argument is as follows. Suppose that at equilibrium, areas $s_0, s_1, s_2 \ldots . s_i \ldots .$ are covered by 0, 1, 2 $\ldots .$ $i \ldots$ layers of absorbed molecules. At equilibrium s_0 must be constant, and any condensation on the bare surface must be balanced by evaporation from the surface covered with only a single layer ; or

$$a_1 p s_0 = b_1 s_1 e^{-E_1/RT}, \qquad \ldots\ldots(17)$$

where p is the vapour pressure and E_1 is the heat evolved per mole of atoms absorbed on to the bare surface ; a_1 and b_1 are condensation and evaporation constants. The area covered by one layer can change in four ways : condensation on the bare surface and evaporation from the area covered by two layers increase it, whilst evaporation from the area covered by one layer and condensation on this layer decrease it. Balancing the decrease and increase gives

$$a_2 p s_1 + b_1 s_1 e^{-E_1/RT} = b_2 s_2 e^{-E/RT} + a_1 p s_0, \qquad \ldots\ldots(18)$$

which, on reference to (17), becomes

$$a_2 p s_1 = b_2 s_2 e^{-E/RT}, \qquad \ldots\ldots(19)$$

where E is the latent heat of condensation, and a_2 and b_2 are condensation and evaporation coefficients for the liquid in bulk. The general equation is

$$a_2 p s_{i-1} = b_2 s_i e^{-E/RT}. \qquad \ldots\ldots(20)$$

The total absorption, v, is given by

$$v = v_m \sum_{i=0}^{\infty} i s_i, \qquad \ldots\ldots(21)$$

where v_m is the volume of gas in a unimolecular layer of the total area of surface, $\sum_{i=0}^{\infty} s_i$. Evaluating (21) and using the condition that v should be infinite when p is equal to p_0, the saturation pressure, gives for the isotherm equation :

$$\frac{v}{v_m} = \frac{\nu p}{(p_0 - p)\{1 + (\nu - 1)p/p_0\}}, \qquad \ldots\ldots(22)$$

where

$$\nu = \frac{a_1 b_2}{b_1 a_2} e^{(E_1 - E)/RT}. \qquad \ldots\ldots(23)$$

The derivation of (22) seems perfectly rigorous, yet it is a curious result to find that a liquid can exert a pressure different from its saturation pressure merely because of the presence of localized sites, or at any rate of sites that are different from those for molecules in the bulk liquid. The underlying statistical implication must be that the interchange of two identical molecules, one in the liquid and one on the localized sites, gives an entropy of interchange and, therefore, a decrease in free energy of the absorbed atoms. But according to current statistical ideas, the interchange of identical atoms, even between unlike sites, does not

give rise to a new configuration, and, therefore, cannot contribute to the free energy. There thus arises a real conflict between the result given by an apparently logical evaporation-condensation argument and that given by current statistical arguments.

Cassie (1945 a) has shown that if an entropy of interchange, or of mixing, for molecules absorbed on the localized sites and in the liquid form is introduced, the isotherm equation (22) is readily obtained in the from

$$N = \frac{B\beta}{(p_0 - p)\{\beta + (1-\beta)p/p_0\}}, \qquad \ldots\ldots(24)$$

where N is the number of molecules absorbed, B is the number of localized sites available, and

$$\frac{1}{\nu} = \beta = \frac{j_L}{j_s} e^{-w/RT} = \frac{b_1 a_2}{a_1 b_2} e^{-(E_1 - E)/RT} = \frac{p_{\frac{1}{2}}}{p_0}, \qquad \ldots\ldots(25)$$

where j^L is the partition function for a molecule in the liquid, j_s is that for a molecule in a localized site, w is the energy evolved when a mole of molecules is transferred from the liquid to localized sites, and $p_{\frac{1}{2}}$ is the pressure required to fill one half of the localized sites in the absence of molecules in the liquid form. The number of molecules absorbed on to the localized sites, X, may also be shown to be given by

$$p/p_0 = (N - X)/N \qquad \ldots\ldots(26)$$

and

$$(N - X)(B - X) = \beta X^2. \qquad \ldots\ldots(27)$$

Brunauer, Emmett and Teller's isotherm has now been widely accepted and used in interpreting experimental isotherms (*cf.* Bangham, 1943), although its statistical implications may not have been realized. The conflict between the statistical ideas that are apparently implied by their equation and the current definition of a new configuration may ultimately be resolved when a better understanding of the statistics of liquids and co-operative phenomena in general has been attained. The abandonment of B.E.T.'s isotherm because of its apparent conflict with current statistics would not be justified until the theory of liquids and co-operative phenomena has been cleared up, and it will be shown in the following sections that it explains very easily both the absorption isotherm of wool and the diffusion of water through keratin.

§ 8. ANALYSIS OF THE WOOL-WATER ISOTHERM

Analysis of textile water-vapour isotherms has followed the ups and downs of the various hypotheses proposed for multimolecular absorption. Swan and Urquhart (1927) give a review of the earlier work. The structure of cellulose and keratin as revealed by chemical and x-ray analyses has now given a considerable degree of faith in the hypothesis that hydrophilic groups, such as CO, OH, and NH_2, associated with the long-chain molecules are primarily responsible for water absorption, and as with multimolecular absorption, discussion now centres on the origin of the deviation of the isotherms from those for monolayers.

Peirce (1929) correlated the cotton isotherm with changes in the physical properties of cotton hairs on absorbing water, and divided the absorbed water into two forms: the (*a*) form, in which one water molecule is definitely associated

with one hexose unit; and the (*b*) form, in which the water molecules fill available spaces under attractive forces like those in a liquid. He used an evaporation-condensation mechanism to develop relations between the water vapour pressure and the amounts of water absorbed in the (*a*) form and in the (*b*) form. Peirce thus gave the first quantitative relationship between the vapour pressure and the amount of the water in the two forms. His method and results resemble those of Brunauer, Emmett and Teller, but he was unable to give the (*b*) water a precise interpretation. He showed that the rigidity of cotton hairs decreased linearly with the amount of (*a*) water absorbed.

Speakman (1944) used Peirce's method in an analysis of the wool-water isotherm. He postulated three forms of absorbed water, one on low-energy localized sites, another on localized sites but of the same internal energy as that of water molecules in bulk, and a third which he termed "capillary" water. The existence of the second type of localized site is, however, very doubtful, because localization of a molecule implies a decrease in the value of its partition function and an increase in its chemical potential, so that equilibrium cannot be maintained between the localized and liquid molecules without a decrease of internal energy on localization.

Bull (1944) has applied Brunauer, Emmett and Teller's (1938) isotherm equation to the absorption of water by proteins, and found a reasonable fit with experiment up to 50% relative humidity. He applied it to the isotherms without correction for hydrostatic stress due to natural restraints, and it is doubtless the absence of this correction that causes serious discrepancies between theory and experiment at higher relative humidities.

Bull evaluated the constants in the isotherm equation, but Brunauer, Emmett and Teller's derivation of it makes interpretation of the constants uncertain. They assume that absorption is always on to a free surface, and when the localized sites are chemical groups of a large molecule in the bulk of a solid, the meaning of the term "free surface" becomes doubtful. Again, Brunauer, Emmett and Teller assume that the evaporation-condensation coefficients are such that the parameter ν of equations (23) and (25) may be taken as $\exp[(E_1-E)/RT]$ without serious error. The statistical analysis shows, however, that the coefficient is $\frac{j_s}{j_L}\exp[(E_1-E)/RT]$, and as j_s must be less than j_L, by the hypothesis of localization, the error in taking j_s/j_L as unity must always be considerable. The values of (E_1-E) derived by Brunauer, Emmett and Teller do, in fact, conflict with experimental values in just the way expected (*cf.* Cassie, 1945 a).

The statistical analysis gives directly a value for B, the number of localized sites per unit mass of solid. Also, equation (26) gives a direct evaluation of the number of localized molecules, X, at any regain, and its value can be checked with measurements of the heat of mixing of N molecules of water with a given weight of dry sorbent; the heat of mixing is, in fact, ωX.

Cassie (1945 b) has used this method to analyse the stress-free isotherm of wool. The analysis is given in table 4. It will be noted that at regains below 0·4 mole, or 7%, the absorbed water is almost exclusively absorbed on to the low-energy localized sites. This must be expected on any analysis that uses the hypothesis of low-energy localized sites, because the free energy of molecules

on these sites will always be small when most of the sites are vacant. This is the deduction that has been used § in 6 to evaluate the coefficient of P in equation (10) by adjusting it to give a constant value for the corrected differential head of wetting, $(q+PV)$, at low regains ; it is seen from table 3 that $(q+PV)$ is, in fact, adjusted to be constant for regains less than 7%.

ΔH in table 4 is the heat of mixing corrected for the swelling energy ; in short, it is $\int_0^m (q+PV)dm$, where m is the regain in moles per 100 gm. of dry wool.

The value of ω deduced from the relation :

$$\Delta H = \omega X \qquad \ldots\ldots(28)$$

is seen to be almost constant at 3500 cal. per mole of water absorbed on to the localized sites from the liquid form.

Another point of interest in table 4 is the value of B, 1·12 moles of sites per 100 gm. of wool. It is tempting to try to identify the sites, but evidence, apart

Table 4

Analysis of the 25° C. stress-free isotherm for 100 gm. wool.

Observed R.H. (%)	N (moles)	Stress-free R.H. (%)	$X^{(1)}$ (moles)	$X^{(2)}$ (moles)	ΔH (k cal.)	ω (k cal./mole)
6·2	0·2	2·1	0·196	0·196	0·6	3·4
22	0·4	5·0	0·380*	0·380*	1·3	3·4
42	0·6	9·1	0·545	0·547	1·9	3·5
63	0·8	14·3	0·685	0·685	2·5	3·6
75	1·0	20·3	0·797*	0·797*	2·9	3·6
79	1·2	26·3	0·885	0·877	3·0	3·4
85	1·4	33·0	0·94	0·94	3·1	3·4
95	1·6	40·0	0·96	0·97	—	—
100	1·8	46·0	0·97	0·99	—	—

* Used to determine B and β : $B=1{\cdot}12$ mole ; $\beta=0{\cdot}10$.
(1) From columns (2) and (3) and equation (26).
(2) From equation (27).

from their number, is scanty. Astbury (1942) in a review of the chemical and x-ray analyses of keratin gives the numbers of amino acid residues present in 100 gm. of wool, and from these data it is possible to derive the number of each of the possible water-attracting groups. The group that most closely approximates to 1·12 moles per 100 gm. is the carbonyl group, which is present to the extent of 1·11 moles. This would be expected to be the most powerful water-attracting group except that much of it occurs in the peptide chain, where it is bonded to imido groups by hydrogen bonds. Pauling (1945) uses Bull's (1944) analysis of the protein-water isotherms to show that there is one localized site per polar amino-acid side chain, and concludes that the carbonyl groups of the main chains do not attract water. There is, however, considerable difficulty in identifying the localized sites in wool with only polar side-chain groups ; the carboxyl side-chain groups can be methylated and the amide groups can be acetylated

(Blackburn and Phillips, 1944) without causing appreciable change in the wool-water isotherm, and if they were the localized sites, a large change in the isotherm should occur. It has, in fact, so far proved impossible to change the wool-water isotherm significantly without destroying the peptide chains and the fibre characteristics of the material.

§ 9. DIFFUSION IN KERATIN

The rate of sorption of molecules by textile fibres is almost as important to the industry as are the equilibrium sorptions. The striking example of water absorption has already been mentioned (§ 5), and although the rate of sorption proved too rapid to be measured, the consequences of the rapid rate of sorption are all important in the use of hygroscopic fibres for clothing. It is, however, in dyeing and in similar wet processes that the manufacturer becomes immediately

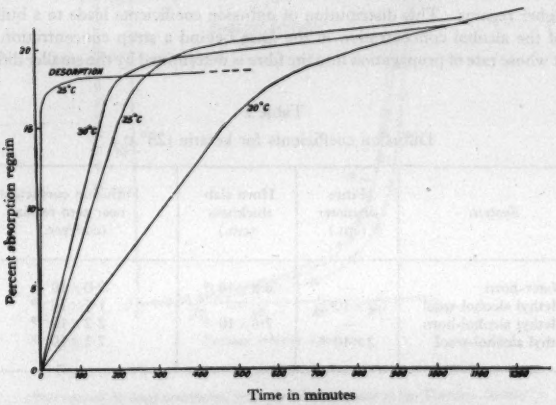


Figure 8. Sorption of ethyl alcohol vapour by wool.

(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

concerned with the rate of sorption of molecules by fibres and the factors controlling the rate of sorption. Innumerable publications have appeared in the technical journals dealing with the rate of exhaustion of dye baths; the work they describe is largely empirical and aims at determining the optimum conditions for rapid and even dyeing of a given fibre with a given dyestuff. Neale and Stringfellow (1933) and Garvie and Neale (1938) have attempted more theoretical investigations of the rate of sorption of dyes by studying the diffusion of direct cotton dyestuffs into regenerated cellulose sheets. They compared their data with Fick's diffusion law, and found that the diffusion coefficient varied with the concentration of dyestuff and the concentrations of salts in the dye bath.

The sorption of molecules from a dye bath, where many different chemical components are present, is not a promising line of attack on the problem of diffusion in textile fibres, and King (1945 a) has recently made a study of the

rate of sorption by wool of alcohols in the vapour phase. Figure 5 shows the temperature-time curves when methyl and ethyl alcohol are absorbed by wool. It is seen that for ethyl alcohol the temperature of the wool does not exert the controlling influence that it did for the absorption of water vapour ; the sorption by the fibres is so slow that the temperature rises by only a negligible amount, and diffusion into the fibre becomes the controlling process. Methyl alcohol occupies an intermediate position where heat and diffusion effects are almost equally important.

Figure 8 shows the absorption of ethyl alcohol by wool plotted against time of exposure to the alcohol vapour. The curves do not fit Fick's diffusion law with a constant diffusion coefficient ; they suggest a constant rate of propagation into the fibre. King has shown that curves of this form would be expected if the diffusion coefficient at low alcohol regains was very small compared with those at higher regains. This distribution of diffusion coefficients leads to a building-up of the alcohol concentration in the fibre behind a steep concentration wave front whose rate of propagation into the fibre is determined by the smaller diffusion

Table 5

Diffusion coefficients for keratin (25° c.)

System	Fibre diameter (cm.)	Horn slab thickness (cm.)	Diffusion coefficient near zero regain ($cm^2/sec.$)
Water-horn	—	$6{\cdot}8\times10^{-3}$	$9{\cdot}0\times10^{-9}$
Methyl alcohol-wool	2×10^{-3}	—	$1{\cdot}7\times10^{-10}$
Methyl alcohol-horn	—	$7{\cdot}6\times10^{-3}$	$2{\cdot}2\times10^{-10}$
Ethyl alcohol-wool	2×10^{-3}	—	$7{\cdot}2\times10^{-12}$

coefficients. King has been able to estimate the smaller diffusion coefficients from the slope of the sorption-time curve. The values are given in table 5. Horn was used to study the diffusion of water vapour because it is also keratin which appears to be physically and chemically very similar to wool, and it can be used in sheets to avoid the control of sorption by the heat effects due to the large surface-volume ratio of wool fibres. The diffusion coefficients determined for methyl alcohol diffusing into wool and into horn are seen to be identical to within the limits of experimental error.

King (1945 b) has extended his study of the diffusion of water in horn by measuring the permeability of horn films to water vapour. The films were sround 5×10^{-3} cm. thick, and were cut parallel to the direction of orientation of the long-chain molecules, so that flow was normal to this direction, which was also the direction of flow for the data of table 5. The films were used as a diaphragm between two compartments, one of which was evacuated, whilst the other contained water vapour at a constant known pressure. The flow through the diaphragm was measured by noting the increase of pressure in the evacuated

chamber for short time intervals. Figure 9 shows the flow of water vapour plotted against the relative humidity on the high-pressure side of the membrane. There are two curves in figure 9, one corresponding with an increasing and the other with a decreasing vapour pressure on the membrane. The hysteresis shown by the two curves is merely a reflection of the regain-relative humidity hysteresis.

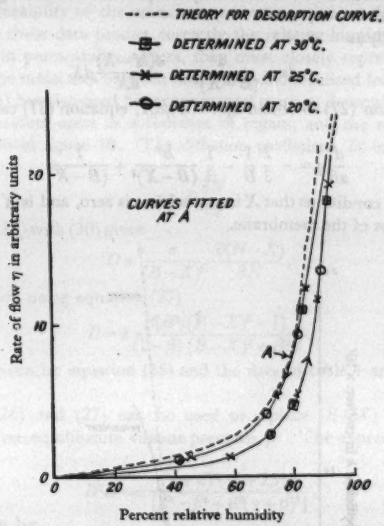


Figure 9. Flow of water vapour through horn keratin membrane.

(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

The curves are characteristic of the permeability of many absorbing solid materials, and Barrer (1941) has suggested that its form is due to the change in activation energy required for hole formation in the lattice as the material swells and changes its elastic constants. King (1945 b) has, however, given a much simpler explanation of the shape of the curve, based on the analysis of water absorption of the previous section.

He assumes that the diffusing molecules are those in the liquid form, and that the concentration gradient of these molecules is $d(N-X)/dx$, where x is the direction of flow. Knudsen's cosine law then gives for the rate of transport, η, across unit area normal to the direction of flow (*cf.* Lennard-Jones, 1932):

$$\eta = \alpha' l^2 \frac{d(N-X)}{dx}, \qquad \ldots\ldots(29)$$

where α' is a constant that may contain an activation-energy term for diffusion of the liquid molecules, and l is the mean free path of the diffusing molecules.

King now assumes that the vacant localized sites act as centres for trapping the mobile molecules, so that l is inversely proportional to $(B-X)$, and (29) becomes

$$\eta = \frac{\alpha}{(B-X)^2} \cdot \frac{d(N-X)}{dx}, \qquad \ldots\ldots(30)$$

where α is independent of the water vapour pressure on the high-pressure side of the membrane.

Rewriting (30) as

$$\eta dx = \frac{\alpha}{(B-X)^2} \cdot \frac{d(N-X)}{dX} dX \qquad \ldots\ldots(31)$$

and using equation (27) to obtain $d(N-X)/dX$, equation (31) can be integrated to give

$$\frac{d}{\alpha\beta}\eta = -\frac{2}{3}\frac{1}{B} - \frac{1}{3}\frac{B^2}{(B-X)^3} + \frac{1}{(B-X)} \qquad \ldots\ldots(32)$$

for the boundary conditions that X is zero when x is zero, and is X when x is equal to d, the thickness of the membrane.

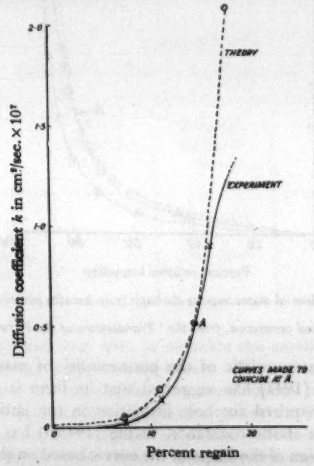


Figure 10. Diffusion coefficient for water in horn keratin.

(*Reproduced, by kind permission, from the 'Transactions of the Faraday Society'.*)

The data of table 4 (§ 8) can be used in equation (32) to evaluate η in arbitrary units. Fitting the values to the experimental curve at 75% relative humidity gives the broken line of figure 9. The agreement of this simple theory with the experimental data is seen to be exceedingly good. The theory predicts a rather more rapid increase of flow than is observed experimentally at higher relative humidities; this discrepancy must be expected because it has been assumed that only vacant localized sites deactivate the diffusing molecules, and

when all the sites are filled, the diffusion coefficient will become infinite, whereas, in fact, other deactivating centres will become important. The theory could be elaborated to include other deactivating centres, but it is clear that at relative humidities up to 80%, the factor determining the flow through the membrance is primarily the number of vacant localized sites.

It is satisfactory that the data deduced from the isotherm analysis explain so well the permeability of the membrane to water vapour at different regains; in particular, as these data predict correctly the relative humidity at which the sudden change in permeability occurs, they must closely represent the correct distribution of the molecules between the mobile and localized forms.

King (1945 b) has also used his experimental work to determine the diffusion coefficient in absolute units as a function of regain, and the relation between the two is shown in figure 10. The diffusion coefficient, D, is defined by the equation

$$\eta = D\frac{\partial N}{\partial x}, \qquad \ldots\ldots(33)$$

and comparing (33) with (30) gives

$$D = \frac{\alpha}{(B-X)^2}\cdot\frac{\partial(N-X)}{\partial N}, \qquad \ldots\ldots(34)$$

which becomes on using equation (27)

$$D = \alpha\frac{\beta[B^2/(B-X)^2-1]}{(1-\beta)(B-X)^2+B^2\beta}. \qquad \ldots\ldots(35)$$

The curve given by equation (35) and the data of table 4 are reproduced in figure 10.

Equations (26) and (27) can be used to replace $(B-X)$ in equation (35) with the stress-free equilibrium vapour pressure, p. The expression for D then becomes

$$B^2D = \alpha\frac{(y+\beta)^2(y^2+2\beta y)}{[\beta^3-\beta^4+\beta^2(y+\beta)^2]}, \qquad \ldots\ldots(36)$$

where y is given by

$$\frac{1}{y} = \left(\frac{p_0}{p}-1\right), \qquad \ldots\ldots(37)$$

p_0 being the saturation water vapour pressure.

The interesting feature of equation (36) is that it shows that any activation energy for D will be dependent on the relative humidity p/p_0. In fact, on differentiating (36) with respect to the temperature and maintaining p/p_0 constant, it is found that if W is the activation energy for p/p_0 zero, then the activation energy when p/p_0 is 0·5 is $(W-1{\cdot}5w)$ for the value of β, 0·1, given in table 4.

King's curves (figure 9) show no activation energy when permeability is plotted against relative humidity, but it is difficult to measure changes of permeability at all accurately at low humidities because of the small gradient of the curves. The temperature coefficient for the diffusion constant at low regains can, however, be determined from rate of sorption measurements, and King (1945 b) has used these to determine the temperature coefficient for water vapour. The activation energy is found to be 7500 cal./mole for the diffusion constant referred to regain, or 4000 cal./mole when referred to relative humidity. There is thus an activation energy of 4000 cal./mole near zero regain and zero activation

energy at high regains. As w is 3500 cal./mole, the decrease of activation energy with increasing relative humidity predicted by the theory is also shown by the experimental data.

A point of practical importance is the magnitude of the diffusion coefficient of water in keratin : it is 1×10^{-7} cm.^{2}/sec. near 15% regain, and may be several times this value near saturation regain, 33%. It seems a small value compared with the diffusion coefficient of water in still air, which is 0·25 cm.^{2}/sec., but it must be remembered that the concentrations of water in keratin are very much greater than those in air in equilibrium with the keratin. Consider, for example, the transport of water per sq. cm. through a slab of keratin 1 cm. thick with air at 20° c. and 90% relative humidity at one side and at 20° c. and 60% R.H. at the other ; the regain at the high-pressure side is 26%, and at the low-pressure side 14%, the corresponding concentrations being 0·34 and 0·18 gm./cm.^{3} for keratin of density 1·3 gm./cm.^{3} If the concentration gradient is assumed to be constant, it is 0·16 gm./cm.^{3}/cm., and a diffusion coefficient of 10^{-7} cm.^{2}/sec. gives a transport of $0{\cdot}16 \times 10^{-7}$ gm./cm.^{2}/sec. The water vapour concentrations in the air, on the other hand, are $14{\cdot}8 \times 10^{-6}$ and $10{\cdot}5 \times 10^{-6}$ gm./cm.^{3} respectively, and the transport through a layer of still air 1-cm. thick is $1{\cdot}1 \times 10^{-6}$ gm./cm.^{2} ; or the transport through still air is roughly 70 times that through the keratin. This calculation has assumed a diffusion constant of 1×10^{-7} cm.^{2}/sec. for keratin, which is the value for transport normal to the molecular direction at 15% regain ; at higher regains and for transport parallel to the molecular direction, the diffusion constant may be much greater. Barrer (1941) quotes values for diffusion along and normal to the planes of laminae of heulandite, and states that the permeability parallel to the planes may be as much as 10^5 times that normal to the planes. It is therefore not inconceivable that the transport of water along wool fibres may be just as great as that through equivalent layers of still air, and that wool fibres in clothing offer no more resistance to the diffusion of water vapour from a perspiring skin than does the still air held in the interstices of the fabric.

§ 10. WATER-REPELLENCY AND TEXTILES

The water-repellency of textiles is a problem of very great practical importance, and to a physicist it should be easily solved, but the complex geometry involved and the variable details encountered in wear have appeared so formidable that few attempts have been made to elucidate the water-repellency—or lack of water-repellency !—of clothing.

Cloths are usually rendered water-repellent by depositing particles of paraffin wax on the fibres and subsequently heating them to spread the wax. It must be remarked, however, that there is no evidence that the heat treatment does spread the wax, but apart from this detail, it is quite certain that the waterproofing processes do provide finite fibre-water contact angles. Clean virgin wool generally has large fibre-water contact angles, and proofing may or may not increase them. Cotton hairs are also water-repellent in their natural state, but their water-repellency is due to a wax surface which is removed in processing, and finished cotton fabrics are generally composed of fibres of zero advancing contact angle. All artificial fibres also appear to have zero advancing contact angles.

Cassie and Baxter (1944) have explained the behaviour of fabrics during their first exposure to rain by extending the theories of Adam (1941) and Wenzel (1936) for the apparent contact angle of rough surfaces to porous surfaces. The cosine of the contact angle, θ, of a surface is a convenient measure of the ratio of the energy gained in forming unit geometrical area of the solid-liquid interface to the energy, γ_{LA}, required to increase the liquid surface by unit area. If the surface is smooth, the usual relation :

$$\cos\theta = \frac{\gamma_{SA} - \gamma_{SL}}{\gamma_{LA}} \qquad \ldots\ldots(38)$$

is obtained where γ_{SA} is the solid-air surface tension, and γ_{SL} is the solid-liquid interfacial tension. If the work gained by the formation of the liquid-solid interface exceeds that required to form the liquid-air interface, the liquid continues to spread—there is no equilibrium and no contact angle. It is the aim of wetting agents to attain this condition. If, however, there is a finite contact angle, and the interface enclosed by unit geometrical area is roughened so that its total area becomes σ, where σ is greater than unity, then the energy gained in forming unit geometrical area of interface is $\sigma(\gamma_{SA} - \gamma_{LS})$, and equation (38) becomes

$$\cos\theta' = \frac{\sigma(\gamma_{SA} - \gamma_{SL})}{\gamma_{LA}} = \sigma\cos\theta, \qquad \ldots\ldots(39)$$

where θ' is the apparent contact angle for the surface. This is Wenzel's result. If θ is less than 90°, roughening the surface will increase $\cos\theta'$, or the apparent contact angle becomes less than the true one ; if θ is greater than 90°, roughening the surface increases the apparent contact angle. It is well known that rough surfaces may give apparently very large contact angles as predicted by the theory, but it must be remembered that close inspection of the surface will always reveal the true contact angle over an area sufficiently small to be smooth.

Wenzel assumes complete contact between the solid surface and the liquid, but on a porous surface complete contact is impossible. Water under no pressure will not pass through the pores so long as the advancing contact angle is finite and the pores are reasonably small, and the interface will now consist partly of areas of solid-water contact and partly of areas of water-air contact. Suppose, then, that f_1 is the total area of solid-liquid interface and f_2 is the total area of liquid-air interface in unit geometrical area parallel to the porous surface over which the water is being spread. The energy gained in forming unit geometrical area of the porous surface-liquid interface is now $f_1(\gamma_{SA} - \gamma_{LS}) - f_2\gamma_{LA}$ and the apparent contact ange, θ'', is given by

$$\cos\theta'' = f_1\cos\theta - f_2, \qquad \ldots\ldots(40)$$

which, of course, reduces to Wenzel's equation when f_2 is zero.

f_2 is always positive, and the presence of pores in a surface will always increase the apparent contact angle. Cassie and Baxter have checked equation (40) by making up a flat cage of fine parallel wires so that it could be used as a plate in Adam and Jessop's (1925) apparatus to measure the contact angle θ'', or rather to measure the tilt of the cage which showed a minimum disturbance of the water surface. f_1 and f_2 were determined from the geometry of the wires and the advancing wire-water contact angle, which was also measured. The

results are shown in table 6, and it is seen that the calculated angles are in very close agreement with the observed values. The wires were coated with paraffin wax and had true advancing and receding contact angles of 105° and 93° respectively. The true advancing contact angle is, of course, always used to calculate f_1 and f_2.

Baxter and Cassie (1945) have applied these methods to textiles, and the agreement with experiment for advancing contact angles is as good as can be expected from the uncertain geometry of fibres in yarns and fabrics. Wool yarns show advancing contact angles as great as 150° to 160°, and are usually around 10° greater than the calculated angles, doubtless due to the low fibre density at the surface of the yarn. Another check can be obtained by measuring the hydrostatic pressure required to force water through a fabric. It can be proved (*cf.* Baxter and Cassie, 1945) that the minimum radius of curvature, R, of a water surface making a contact angle θ_A with parallel cylinders of radius r, with their axes $2(r+d)$ apart, is given by

$$R = r\cos\theta_A + \sqrt{(r+d)^2 - r^2\sin\theta_A}, \qquad \ldots\ldots(41)$$

and the pressure required to force water through the square pores between the

Table 6

Apparent contact angles of porous surfaces

Wire diam. (cm.)	Wire spacing (cm.)	f_1	f_2	θ''_A calc.	θ''_A obsd.	θ''_R calc.	θ''_R observed
				°	°	°	°
0·007	0·0250	0·366	0·730	145·5	143·0	138·5	138·0
0·007	0·0404	0·227	0·833	153·0	152·0	147·5	148·0
0·013	0·0250	0·680	0·498	132·5	132·5	122·0	118·0
0·013	0·0404	0·422	0·689	143·0	143·0	135·5	135·5

warp and weft yarns of a fabric must be given by $2\gamma_{LA}/R$. The appropriate advancing contact angle is, of course, the apparent advancing contact angle for the yarns. Using this angle, Baxter and Cassie (1945) have shown that the pressure required to force water through a fabric should be 27 cm. of water, whilst the measured value was 25 cm. In this case the calculated yarn advancing contact angle was 150°, and the fibre angle was 120° ; the latter when used in equation (4) gives the hydrostatic pressure as 18 cm. This example shows that the porous structure of yarns gives a real advantage in increasing the pressure required to force water between the yarns by as much as 50% over that given by solid "yarns" with the fibre-water contact angle. It should be remarked that the large fibre-water advancing contact angle, 120°, quoted in this example is quite common for wool fibres ; it is due to the scales providing a rough surface, and, therefore, an apparently large advancing contact angle.

The position regarding receding angles for textiles is not so satisfactory. When a yarn is first immersed in water and withdrawn, its receding contact angle is large, but it rapidly decreases to a value much below that predicted by theory ;

thus, where the calculated angle is 100°, the angle finally observed may be no more than 20°. The immediate cause of this discrepancy is obvious from an examination of the yarns after prolonged immersion: water films form between the fibres, and as the fibres in a yarn are not parallel, they have points of intersection and form closed boundaries around the films. Calculation of the energy gained by wetting out the peripheries of the fibres shows that it is not generally large enough to compensate for the energy expended in forming the water-air interfaces. The presence of these films reduces the contact angle to zero locally and so gives the observed small receding contact angle; but the correct explanation of the formation of the film is not at all clear.

An account of the behaviour of a fabric in rain can be given in terms of the phenomena which have just been discussed. A raindrop falling on to a dry fabric will have large apparent advancing and receding contact angles; it will form itself into a drop and roll off the fabric. After prolonged exposure, the receding angle decreases, due to film formation, the drop no longer forms a sphere and the fabric becomes wet. If film formation could be avoided, the fabric would never become wet. The duck's feather appears to have a structure that makes film formation difficult. A feather consists of a main stem or rachis, which carries barbs on either side; the barbs are visible to the unaided eye and give the feather its characteristic appearance. Each barb supports fine fibres known as barbules, on either side of it. The barbules on one side of the barb carry hooks and those on the other side carry notches; the two types of barbules from neighbouring barbs overlap, and the hooks from one set engage the notches of the other, so that the outer surface of the feather consists of fine parallel fibres that cannot be drawn together by surface tension forces, and the formation of closed boundaries is avoided. The barbules of breast feathers are roughly 8μ in diameter, and they are spaced almost 40μ apart, so that the surface of the feather has a very large porosity. The advancing and receding contact angles of the solid rachis are around 100° and 90° respectively, and equation (40) gives the apparent angles as 150° and 130°. The large contact angles of the duck's feather are thus due to its physical structure rather than to any exceptional proofing agent.

Nature appears to make extensive use of the physical structure of surfaces for water-repellency. The fibres of an animal's fur are pulled together by surface tension at the tips to wet out in the characteristic tufts, but near the skin the fibres are held apart—even in the sheep only 2% of the skin area consists of hair follicles—and a large apparent advancing contact angle prevents wetting near the skin. Many leaves (*cf.* Woog, 1942) have rough cuticles, and the total reflection of the "silver" drops so often to be seen on leaves, reveals the tiny contact that water makes with the solid substance of the cuticle. The apparent contact angles are very large, and the diurnal changes in the angles observed by Fogg (1944) are more likely to be due to changes in the structure of the cuticle than to changes in its composition (*cf.* Cassie and Baxter, 1945).

ACKNOWLEDGMENTS

The author thanks the Director of Research and the Council of the Wool Industries Research Association for permission to contribute this *Report*.

REFERENCES

ADAM, N. K., 1941. *Physics and Chemistry of Surfaces* (Oxford : University Press), 3rd Ed. p. 186.

ADAM, N. K. and JESSOP, G., 1925. *J. Chem. Soc.*, 1865.

ANGUS, T. C., 1936. *J. Text Inst., Manchr.*, **27**, T. 273.

ASTBURY, W. T., 1942. *J. Chem. Soc.*, 339.

BABBITT, J. D., 1942. *Canad. J. Res.*, **20**, 143.

BANGHAM, D. H., 1943. *Ann. Rep. on Progress of Chemistry*, **40**, 38.

BARKAS, W. W., 1942 a. *Trans. Faraday Soc.*, **38**, 194.

BARKAS, W. W., 1942 b. *Trans. Faraday Soc.*, **38**, 447.

BARKAS, W. W., 1945. *Swelling Stresses in Gels* (London : H.M. Stationery Office), F.P.R. Special Report No. 6.

BARKER, S. G., 1929. *Wool* (London : H.M. Stationery Office ; Empire Marketing Board), **21**, 62.

BARRER, R. M., 1941. *Diffusion in and through Solids* (Cambridge : University Press), p. 102.

BAXTER, S., 1946. *Proc. Phys. Soc.*, **58**, 105.

BAXTER, S. and CASSIE, A. B. D., 1941. *Trans. Faraday Soc.*, **37**, 517.

BAXTER, S. and CASSIE, A. B. D., 1943. *J. Text. Inst., Manchr.*, **34**, T. 41.

BAXTER, S. and CASSIE, A. B. D., 1945. *J. Text. Inst., Manchr.*, **36**, T. 67.

BLACK, C. P. and MATTHEW, J. A., 1934. *J. Text. Inst., Manchr.*, **25**, T. 197.

BLACKBURN, S. and PHILLIPS, H., 1944. *Biochem. J.*, **38**, 171.

DE BOER, J. H. and ZWIKKER, C., 1929. *Z. Phys. Chem.*, **133**, 407.

BRADLEY, R. S., 1936. *J. Chem. Soc.*, 1467.

BRITISH STANDARDS INSTITUTION, 1942. No. 1051.

BRUNAUER, S., 1942. *The Adsorption of Gases* (Oxford : University Press), **1**.

BRUNAUER, S. and EMMETT, P. H., 1937. *J. Amer. Chem. Soc.*, **59**, 2682.

BRUNAUER, S., EMMETT, P. H. and TELLER, E., 1938. *J. Amer. Chem. Soc.*, **60**, 309.

BULL, H. B., 1944. *J. Amer. Chem. Soc.*, **66**, 1499.

CASSIE, A. B. D., 1940 a. *J. Text. Inst., Manchr.*, **31**, T. 17.

CASSIE, A. B. D., 1940 b. *Trans. Faraday Soc.*, **36**, 453.

CASSIE, A. B. D., 1945 a. *Trans. Faraday Soc.*, **41**, 450.

CASSIE, A. B. D., 1945 b. *Trans. Faraday Soc.*, **41**, 458.

CASSIE, A. B. D., ATKINS, B. E. and KING, G., 1939. *Nature, Lond.*, **143**, 163.

CASSIE, A. B. D. and BAXTER, S., 1940. *Trans. Faraday Soc.*, **36**, 458.

CASSIE, A. B. D. and BAXTER, S., 1944. *Trans. Faraday Soc.*, **40**, 546.

CASSIE, A. B. D. and BAXTER, S., 1945. *Nature, Lond.*, **155**, 21.

DANIELS, H. E., 1941. *Trans. Faraday Soc.*, **37**, 506.

DENHAM, W. S. and ALLEN, A. L., 1933. *Trans. Faraday Soc.*, **29**, 316.

DIETZ, O., 1912. *Wochense. Papierfab.*, **43**, 3119.

FISHER, E. A., 1923. *Proc. Roy. Soc.*, A, **103**, 139.

FOGG, G. E., 1944. *Nature, Lond.*, **154**, 515.

FORBES, G., 1872–3. *Proc. Roy. Soc., Edinb.*, **8**, 66.

GARVIE, W. M. and NEALE, S. M., 1938. *Trans. Faraday Soc.*, **34**, 335.

HEDGES, J. J., 1926. *Trans. Faraday Soc.*, **22**, 178.

HENRY, P. S. H., 1939. *Proc. Roy. Soc.*, A, **171**, 215.

HERCUS, E. O. and SUTHERLAND, D. M., 1934. *Proc. Roy. Soc.*, A, **145**, 599.

HILL, A. V., 1928. *Proc. Roy. Soc.*, B, **104**, 71.

KATZ, J. R., 1917–18. *Kolloidchem. Beitr.*, **9**, 1.

KING, A. T., 1926. *J. Text. Inst.*, **17**, T. 53.

KING, G., 1945 a. *Trans. Faraday Soc.*, **41**, 325.

KING, G., 1945 b. *Trans. Faraday Soc.*, **41**, 479.

KING, G. and CASSIE, A. B. D., 1940. *Trans. Faraday Soc.*, **36**, 445.

LEES, C. H., 1900. *Phil. Mag.*, **49**, 221.

LENNARD-JONES, J. E., 1932. *Trans. Faraday Soc.*, **28**, 349.

MARSH, M. C., 1931. *J. Text. Inst., Manchr.*, **22**, T. 245.

MATTHEWS, J. M., 1923. *Textile Fibres* (New York : Wiley and Sons).

NEALE, S. M. and STRINGFELLOW, W. A., 1933. *Trans. Faraday Soc.*, **29**, 1168.

NEALE, S. M. and STRINGFELLOW, W. A., 1941. *Trans. Faraday Soc.*, **37**, 525.
NIVEN, C. D. and BABBITT, J. D., 1938. *J. Text. Inst., Manchr.*, **29**, T. 161.
PAULING, L., 1945. *J. Amer. Chem. Soc.*, **67**, 555.
PEIRCE, F. T., 1929. *J. Text. Inst., Manchr.*, **20**, T. 133.
PEIRCE, F. T., 1930. *J. Text. Inst., Manchr.*, **21**, T. 377.
PRESTON, T., 1929. *Heat* (London : Macmillan), 3rd Ed., p. 506.
REES, W. H., 1941. *J. Text. Inst., Manchr.*, **32**, T. 149.
RUBNER, M., 1895. *Archiv. f. Hygiene*, **24**, 265.
SCHLOESING, T., 1893. *Bull. Soc. Encour. Indust. Nat.*
SCHUHMEISTER, J., 1877. *Ber. Akad. Wien.* (*Math. Klasse*), **76** (2), 283.
SHORTER, S. A., 1924. *J. Text. Inst., Manchr.*, **15**, T. 328.
SPEAKMAN, J. B., 1927. *J. Text. Inst., Manchr.*, **18**, T. 446.
SPEAKMAN, J. B., 1936 a. *J. Text. Inst., Manchr.*, **27**, T. 185.
SPEAKMAN, J. B., 1936 b. *J. Text. Inst., Manchr.*, **27**, T. 446.
SPEAKMAN J. B., 1944. *Trans. Faraday Soc.*, **40**, 6.
SPEAKMAN, J. B. and CHAMBERLAIN, N. H., 1930. *J. Text. Inst., Manchr.*, **21**, T. 29.
SWAN, E. and URQUHART, A. R., 1927. *J. Phys. Chem.*, **31**, 251.
TROUTON, F. T. and POOL, B., 1906. *Proc. Roy. Soc.*, A, **77**, 292.
URQUHART, A. R. and WILLIAMS, A. M., 1924. *J. Text. Inst., Manchr.*, **15**, T. 559.
WENZEL, R. N., 1936. *Industr. Engng. Chem.*, **28**, 988.
WHIPPLE, F. J. W., 1933. *Proc. Phys. Soc.*, **45**, 307.
WOOG, P., 1942. *Arch. Phys. Biol. Chem.-Phys. Corps. Organ.*, **16** (53), 15.

SOME APPLICATIONS OF PHYSICS TO THE PROCESSING OF TEXTILES

By J. G. MARTINDALE,
Wool Industries Research Association

§1. INTRODUCTION

The textile industries provide great scope for the application of physics and physical methods of investigation to the examination of their many types of raw materials, semi-manufactured and finished products and the physical and chemical processes through which textile materials pass in the course of manufacture.

This will be appreciated readily by the reader even though he may not have any detailed knowledge of the requirements of either textile materials or processes. If one merely considers the large number of finished textiles which are everyday articles of personal, domestic or industrial equipment, and the very different kinds of service which each is called upon to give, the diversity of the problems presented is sufficiently striking. By way of examples one could quote conveyor belts and the problem of their flexibility and durability, curtain materials and their fastness to light and to washing, the warmth-retaining, draping and crease-resisting qualities of articles of clothing, the felting of woollen underclothing, the lustre of mercerized cotton stockings, the electrical insulating and fire-resisting properties of textile yarns and fabrics, the strength of an ordinary piece of string or sewing cotton, the waterproofing of fabrics and many other equally diverse examples. Add to this the properties of the great variety of intermediate products of manufacture, the behaviour of the material in its passage through the machines, the multiplicity of types of machines employed, and the many assistants or dressings used during manufacture—oils, soaps, sizes, resins, etc.—and some idea can be formed of the great variety of the problems which confront the physicists engaged in these fields of investigation.

In this *Report*, one of the first two which have been called for regarding the part that physics has been able to play in textile research, it has been thought desirable to call attention to this variety, however inadequately, and to record that workers are engaged in the widely different pursuits that it entails. The main purpose of the *Report*, however, is to enlarge on one small section of these available fields of research, that is, to describe the progress which has been made in investigating (i) those properties of wool fibres which are of direct importance in the spinning of yarns, and (ii) the behaviour of fibres during the processes leading up to the production of a yarn. It will be for others, on future occasions, to present other aspects of the application of physics to other branches of textiles.

It has been said (von Bergen, 1939) that "we have to admit our ignorance, from the scientific point of view, of many of the essential fundamentals underlying commercial practice in woollen manufacturing". This is equally true of other

branches of textile manufacture. It may, perhaps, not be out of place to remark once again that the textile industries as a whole have hitherto depended for their technical efficiency on the ability of skilled craftsmen, who have been trained in the traditional methods of their predecessors ; and for their technical development on individuals amongst them who have had the ability to examine their experiences critically, and incorporate the results of these observations into their normal technical practices. In this way empirical techniques have gradually evolved. There is a noticeable absence of any written technology, based on well-authenticated experimental evidence, to account for the existence of, or need for, many of the industrial practices commonly adopted. This obscurity has, of course, been aggravated by the secrecy observed by many craftsmen jealous of their rivals. The greater part of any experimental work, which must have been done before the advent of disinterested observers, is unrecorded.

Those physicists who have, in more recent times, claimed acquaintance with the textile industry can, therefore, be said to have come into the same inheritance. Their reaction has generally been one of dissatisfaction with the empirical methods of its procedures, and a desire to search for the fundamental reasons which would afford some explanation of them, and help to establish textile technology on a more satisfying scientific foundation. It is nevertheless true that, where it has been found possible to elucidate the reasons for traditional methods, it is generally found that there is a sound basis for good and accepted commercial practice. It seems likely, therefore, that the greatest progress will be made by the logical development of fundamentals brought to light by such investigations, resulting in the more efficient attainment of the same ends, rather than by any revolutionary discovery that present practices are in some way unsound.

This *Report* attempts to record the methods adopted by physicists who have tried to reach this understanding of the processing of wool, and some of the results so far achieved. In many respects progress may not appear to be very marked, or very rapid, and where the writer can be excused for overlooking relevant information, of which he may be unaware, this can be explained by the fact that, by comparison with other fields of research work, the number of physicists engaged upon these tasks has been small.

§ 2. THE CONVERSION OF WOOL FIBRES INTO WORSTED YARN

In order that the significance and purpose of the work to be described may be appreciated by a physicist, it is necessary to give a simple description of the processes by which wool is converted into a worsted yarn ; this description is by no means complete, and may seem lacking in detail to anyone who is at all conversant with the wool industry ; it is merely intended to give sufficient information to enable the principles of operation to be grasped.

It is also necessary before proceeding further to define a number of technical terms which inevitably arise in the descriptions of research work given later.

There are two very distinct and different processes by which wool is converted into yarns, known as the woollen and the worsted processes ; it is the second of these which is considered in this *Report*. Even in the worsted process there are three or four alternative methods of manufacture, but the principles underlying all are the same.

The following scheme shows the worsted process in rough outline :—

Process	*Object*	*Product*
Shearing	The removal of wool from sheep.	Fleeces.
Classing	The rough classification of fleeces into batches of similar appearance and character.	Classers' lots.
Sorting	The disintegration of the fleeces of one lot into a number of " sorts " or " qualities ". In each of these the fibres are judged by the sorter to be capable of a certain specified and uniform manufacturing performance.	Matchings or qualities (described by a wool quality scale).
Scouring	The removal of grease, yolk, dirt, and some vegetable matter.	Clean scoured wool.
Drying	To dry the wet scoured wool to a desirable moisture content.	
Carding	The further removal of solid impurities, the opening out of the scoured locks of wool into a lofty mass of fibres and the conversion of this into a continuous rope-like strand.	Card sliver.
Gilling*	To straighten the fibres in the card sliver and to orient them, to some extent, in the direction of the sliver axis.	Gilled sliver.
Combing	To remove the last traces of solid impurity, to remove short fibres, to straighten the remainder and orient them very completely in the direction of the sliver axis.	Combed sliver.
Gilling	The blending of several combed slivers to produce a sliver of uniform and standard commercial thickness.	Top.
Grading	The re-examination of the fibres of the tops for the benefit of the subsequent manufacturers. The tops are appraised and regraded, being given quality numbers describing their expected performance and quality in the subsquent processes and products.	A scale of qualities of wool in top form.

The above operations, with the exception of shearing and classing, belong to the branch of industry known as wool-combing. At this point the industry is horizontally divided ; the remaining operations belong to worsted spinning.

Process	*Object*	*Product*
Gilling	The blending of tops of similar qualities and some reduction of the thickness of the rope-like sliver.	Sliver
Drawing	The further gradual attenuation of slivers towards the ultimate thickness of the required yarn. (The name of the attenuation process is changed to drawing because of a change in the type of machine used.)	Slubbings and rovings of various thicknesses.
Spinning	The last process of attenuation of thickness ; it produces a yarn of the required diameter and inserts the necessary turns of twist per unit length of yarn to give it the requisite strength.	Yarn.

A few comments amplifying the details of this brief description of the process may be useful.

No two fleeces, even from the same breed of sheep running in the same flock, are exactly alike, and some may be very different from others, hence the need for a

* The *g* is hard.

rough preliminary classing into similar lots for the purposes of marketing. Furthermore, any one fleece contains very different *qualities* of fibres, fibres in one section of the fleece may be markedly different from those in another section. The wool sorter inspects each fleece individually and splits it up by hand into a number of sorts or qualities. Those fractions from different fleeces which he deems to be of the same quality are put together for future processing. The basis of the sorter's judgment of quality will be discussed later ; it only needs to be said here that the sorting is necessary, otherwise the quality of the final product, and the efficiency with which the conversion could take place, would be adversely affected by the presence of the poorest fibres of the fleeces.

After sorting the wool still retains its natural grease, yolk (mostly potassium salts exuded by the sweat glands) and various deleterious substances such as dirt and vegetable matter. The amount varies with the breed of sheep, but a 50% wool content is quite normal, and values as low as 30% are not unknown. The grease and yolk are removed by scouring in warm alkaline soap solutions, and the more solid adulterants are gradually removed throughout all the operations up to and including combing.

Carding is essentially an attempt to disentangle the locks of wool, which are felted to some slight extent as a result of the wet treatment of scouring, and to open them out into such a tenuous state that each fibre is completely separated from its neighbour. It is then relatively simple to produce a continuous sliver of fibres. This disentanglement is accomplished by a teasing action in a carding machine by repeatedly transferring the wool from slowly-moving to fast-moving cylinders. The surfaces of these cylinders are covered with a clothing of fine, short, metal teeth or pins to which the fibres, or tufts of fibres, attach themselves, and are teased out and carried forward through the carding machine. A weighing hopper is included in the mechanism for the purpose of producing an even flow of wool to the machine and a consequent uniformity of thickness of the emerging sliver.

The interplay of wool between the various cylinders, which present alternative paths for the wool to follow at several points on the machine, offers an interesting field for experiment. A recent paper discussing the distribution and movement of wool on carding machines (Martindale, 1945 b) describes a series of tests in which the probability of fibres following the different courses through the machine was determined experimentally. From this the degree of mixing of the fibres and "levelling" of the flow of wool through a card can be calculated, and it was shown that this is conditioned by the relative speeds and relative settings of the various cylinders.

The wool emerges from the card as a thin web of fibres about 5 ft. wide, which is passed through a narrow funnel and consolidated into an untwisted riband of approximately circular cross-section about 1·5 in. in diameter. This is the *card sliver*.

It is characteristic of a card sliver that the fibres comprising it are neither straight nor parallel. They are present in a somewhat entangled state, although there is a slight orientation in the direction of the sliver axis. The further processing operations in making a worsted yarn are directed towards straightening the fibres, laying them parallel to one another and to the axis of the sliver, removing

the short fibres and reducing the diameter from 1·5 in. to a uniform diameter between 0·025 in. and 0·007 in. according to the type of wool used and the thickness of yarn required.

When fibres exist in a somewhat entangled state in the rope-like form of a card sliver they can be straightened in either of two ways. If the sliver is passed through the two pairs of parallel rollers (figure 1) so spaced that the two pairs cannot both grip the same fibre at the same time, and if the second pair—the delivery rollers—have a greater surface speed than the first pair—the feed rollers—then, when a fibre is gripped by the delivery rollers, its velocity is suddenly increased, and it is drawn through the tuft of fibres still under the control of the feed rollers. This straightens it out and lays it in the direction of movement, which is parallel to the axis of the sliver. The relative surface speeds of the two sets of rollers is known as the *draft*, and the sliver is attenuated in thickness in accordance with the draft applied. The draft is kept below a certain limit so that the supply of fibres to the delivery rollers is maintained, consequently the continuity of the attenuated emerging sliver is preserved.

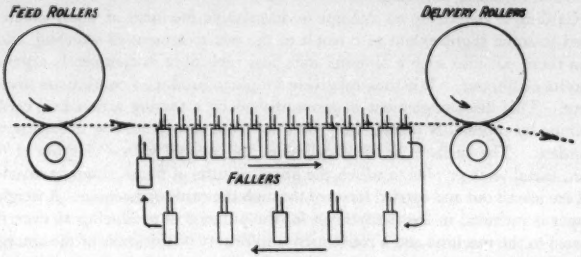


Figure 1.

It is this action which is the basis of all the gilling and drawing operations referred to in the tabulated outline of the processes. By this means fibres are straightened in the slivers and thick slivers are reduced to finer slivers and eventually to yarn.

In the gilling operations assistance is given to the straightening operation by a series of metal combs which penetrate the sliver between the two sets of rollers. These combs are known as *fallers* because they move forward with the fibres at a speed intermediate between that of the delivery and feed roller surfaces, and fall out one by one as they reach the delivery rollers. They are then returned to the back, where they are lifted, and penetrate the sliver again. The study of the relative movements of fibres during this operation of drafting is one that will be dealt with in some detail in the later stages of this *Report*. It is of great importance, because upon it depends the uniformity of thickness of the yarns made, and this in turn affects the strength and other desirable properties of a yarn.

The second method of straightening the fibres in slivers is adopted during combing. In this operation, which is entirely mechanical and automatic, the sliver entering the combing machine is taken to pieces and reduced to a series of fibre tufts, which are combed individually at both ends by fine metal combs. The

combs remove short fibres and any remaining dirt or vegetable matter, and at the same time straighten the fibres in the tufts, which are then reunited to form the continuous combed sliver that emerges.

Combed slivers are built up into thicker tops by gilling several together and combining them, so improving their uniformity. The tops produced are the raw material of the spinner, who grades them according to their quality in the same way as the wool sorter originally judged the greasy wool, that is, he judges them in accordance with what his training and experience dictates will be their processing efficiency during their conversion into yarn.

The conversion of tops to yarn is mainly a question of the attenuation of thickness, the chief attention being paid to the preservation of uniformity of thickness. Uniformity of thickness is affected by the rate of attenuation ; the operation has to be performed gradually in several stages if the desirable attribute of yarn uniformity is not to be sacrificed. In the early stages this reduction is performed by gilling operations, but there comes a point when the sliver is so thin that it will not withstand any further mechanical handling without being damaged, and at this point the sliver must be twisted as it emerges from the delivery rollers to give it strength. As soon as the sliver is twisted it has to be wound on to a bobbin, and the use of the gill boxes using metal faller combs is precluded, as the teeth of these combs will not penetrate a twisted sliver. They are therefore dispensed with and the process of attenuation proceeds without them. It might be mentioned that the amount of twist imparted to sliver has to be increased gradually as the thickness decreases, but it must never be so great as to restrict the easy withdrawal of fibres by the delivery rollers.

Finally, when the sliver has been reduced to the required thickness, it is twisted with a comparatively high degree of twist to bind the fibres securely together and give a yarn which has considerable strength.

It should now be clear that the process of making a yarn is one that can provide the physicist with ample scope for interesting and useful researches, and this is submitted as the reason for describing the industrial methods at such length in this *Report*. Some of the research work which the problems involved have already stimulated is described in the following pages.

§3. THE QUALITY OF WOOL AND THE PHYSICAL PROPERTIES OF ITS FIBRES

(i) *The appraisement of wool quality*

The important part which hairs of all kinds have played in clothing humanity has meant that naturally occurring fibres have for long been the objects of interest and curiosity to scientists. As early as 1665 Hooke was describing the appearance of wool and hair in his book *Micrographia of some Physiological Descriptions of Minute Bodies made by Magnifying Glasses, with observations and enquiries thereupon*. Since then the detailed examination of wool fibres has continued, and a very complete summary of all published work from the time of Hooke has been given by Barker (1931) in his book *Wool Quality*. The publication of this book might be regarded as one of the notable stages of progress in the application of physics to textiles, as for the first time it brought together in one volume descriptions of all notable experimental work that had been carried out on wool up to that time.

If one might single out one person whose work, recorded by Barker, was outstanding, the choice falls on Daubenton (1779). He seems to have been the first to mention wool-fineness measurements and made a big step forward—he measured wools of various types and noted that they differed in average fineness ; he even went so far as to classify wools into several grades of fineness and to correlate fineness with manufacturing properties. He was therefore a pioneer in the attempt to explain the spinning quality of a wool in terms of one of its physical properties. In this he was the instigator of a type of work which still arouses a good deal of interest, and, indeed, at times a good deal of controversy.

In order to understand what is meant by spinning quality it is necessary to say a little more about the spinning operation. For a wool to be suitable for manufacture into a yarn of specified fineness, it must possess the ability to spin efficiently to that fineness, that is to say, without too high a rate of breakdown of the yarns issuing from the delivery rollers of the spinning machine. These yarns, as they pass forward, are highly twisted by a spindle and wound on to a spool carried on the spindle. In this operation tension is applied to the yarn, which must therefore have an inherent minimum strength sufficient to enable it to withstand the stresses of twisting and winding-on without breaking down too frequently. This ability is termed the ***spinning power*** or ***spinning quality*** of a wool, and it is usually expressed in terms of the finest yarn to which the wool will spin efficiently. It is this "quality", and no other, that a wool sorter or grader of tops attempts to estimate when he sorts wool or judges tops.

The fineness of yarns is expressed by a scale of numbers, the finer the yarn the higher its number. By an association of ideas, therefore, the commercial quality scale for wool or tops—in this country—is also a numerical scale, running from 30 to 100, the higher numbers referring to wools with the greater spinning power. Such a scale is necessary for the purpose of standardization and commercial description.

By experience, and long association with his materials, the wool sorter or appraiser of tops is able to assess the spinning quality of a sample of wool fibres. The exact basis of his judgment is not at all clear, however, for he uses both sight and touch in forming his conclusions, yet cannot show precisely how these conclusions are synthesized from his various observations. Moreover, as the process is entirely subjective, as differences of opinion are not uncommon, and as there is sometimes a difference between the appraiser's assessment and the performance of the material during spinning, it has been the physicist's endeavour to find some more objective method of determining fibre quality.

There are two similar lines of approach which these attempts can follow : first, to determine the part played by different physical properties of the fibres in influencing spinning power as found by experiment, and, secondly, to assume that the wool appraiser can give a reliable opinion, which he undoubtedly can do on the whole, and to try to explain the way in which he intuitively grades fibre in terms of their measurable properties.

Apart from the inherent interest in this problem, it would be of great practical importance to solve it from the point of view of both the yarn spinner and the sheep breeder ; to the former in that he would be able to stipulate those fibre

characters which would be most useful to him, and to the latter in that he could conduct his breeding trials to provide them.

Although there are probably a number of physical fibre properties which affect the spinning power of a wool, all are agreed that fineness is the character of paramount importance. Other characters which have been said to influence it are fibre length, shape—both longitudinal shape and shape of cross-section—surface structure, elastic properties and strength. The electrical properties of wool also affect its spinning efficiency. These properties of wool fibres will therefore be considered in turn.

(ii) *Wool fineness and quality*

There are many sensations of both sight and touch which influence the wool-sorter in his classification of wools, but it is also a matter of common knowledge that the fibre fineness is the main character which he takes into consideration. Work by Barker and Winson (1931), on the relation between the appraisers' quality numbers and fibre fineness, has given interesting results. They examined tops, which had been classified by wool appraisers of several countries, and they found that a linear relationship existed between the logarithm of the mean fineness of a wool, expressed as weight per unit fibre length, and the position assigned to it in the scale of qualities. This was true of fibre fineness in the British, French, German and Italian systems of grading. It was pointed out that this was in accordance with the Fechner-Weber law (Fechner, 1860), which may be expressed by the formula $I = C \log S$, where, in general terms, S represents the stimulus and I the sensation, C being constant. Barker and Winson say, regarding the assumption of this law in the case of wool grading : "If we regard the wool sorter's judgment as indicative of I, then it must immediately follow that any attempt on his part to form a gradation of fineness will result in a scale in which successive finenesses increase in geometrical progression. This is exactly the case in practice in all countries."

In addition to the fact that the fibres in a top must possess a certain average fineness in order that it may "make the grade", it is equally well known, to sorters and spinners alike, that there must be some limitation upon the amount of variation of fibre fineness within a *quality*. Both these facts have recently been recognized by the U.S. Government, which on 1 January 1940 promulgated a series of standard fibre requirements for wool tops (A.S.T.M., 1944). This is the first instance of commercial wool-quality standards being defined by any public authority in terms of fibre characters determined by stipulated physical methods. The quality of the wool is defined in terms both of the mean fibre diameter and the variation of diameter. As the quality number increases, the fibre diameter becomes smaller ; for example, a wool classed as 50's quality must have a mean fibre diameter between 31·5 μ and 29 μ, it must not have more than 4% of fibres with diameters between 10 and 20 μ and must not contain more than 1·25% above 50μ diameter. On the other hand, a wool classed as 80's quality has a mean diameter of from 18 to 19·5 μ and must not contain more than 0·25% with diameters between 30 and 40 μ.

It will be seen that both these attempts to relate fineness to quality correlate wool fineness with the sorter's judgment, and no mention has been made of any

attempt to relate fineness directly with the ultimate index of quality—the thickness of the finest yarn which it will be possible to produce from it in actual practice.

As far as the writer is aware, no work has been published on this important aspect. There are reasons, however, for this absence of information, one being the absence of a reliable technique for establishing the spinning limit of a wool. As the fineness of yarn being spun is increased there is no clearly defined critical point beyond which spinning is impossible, it merely becomes a matter of increasing difficulty, the number of yarn breakages in a given time gradually increasing as the yarn becomes finer. This efficiency of spinning also depends on many other variables, such as atmospheric conditions, and the effect of such factors has to be determined before one can eliminate the part which they play. Progress has been made at the Wool Industries Research Association in this problem of designing a spinning-test technique, and results have been obtained showing the way in which the spinning limit increases as fibre diameter decreases.

(iii) *Wool-fineness measurement*

To the physicist involved in such investigations, the need for the accurate measurement of the various fibre properties involved is evident, and in this connection a proper sampling technique is of first importance as all fibre properties are very variable. Statistical implications have therefore to be taken into account in deciding upon any technique of measurement, as well as in considering the significance of the results of the measurement. Although this was realized prior to 1930, it is only since that date that statistical methods have been carefully considered and properly applied in wool-fibre measurement.

The application of statistical methods to the determination of fibre fineness was very well summarized by Stanbury and Daniels (1937). In discussing it they said : "The purpose of a sample is to provide information about the main bulk of material whose exact properties it is impossible to ascertain directly. It is now recognized that there are two distinct types of sample which can be extracted from any longitudinal association of fibres of varying lengths. In one type the proportions of fibres of each length group present are the same as in the bulk of wool, within the limits of random sampling error. Such a sample is called a "true sample", and can be obtained by extracting all fibres which *end* in the volume enclosed by two cross-sections of a sliver separated by a distance less than that of the shortest fibre present." A method of obtaining such a sample has been described by Daniels (1942), and will be described later when discussing the measurement of fibre length.

In the second type of sample, referred to by Stanbury and Daniels (1937) and by Townend (1935, 38), the chance of a fibre being present in the sample is proportional to its length, and this type of sample is obtained by removing all fibres which *cross* a section in any sliver. Such a sample is called a ***length-biassed*** sample.

The effect of such a difference between the two types of sample is most marked of course when the sample property to be measured is mean fibre length, or the variability of length of fibres, but it cannot be ignored if the property being

measured is correlated with fibre length to any extent. Stanbury and Daniels (1937) have shown that fibre length and fibre diameter may be correlated (correlation coefficient $r=0{\cdot}5$), the longer fibres tending to be coarser. Thus a length-biassed sample would give a higher value of fibre diameter than a true sample because it contains a greater proportion of long fibres. They show that if D_l is the mean diameter of a length-biassed sample, and D_t that of a true sample, $D_l=D_t+r\sigma_l\sigma_d/L_t$, where L_t is the true mean fibre-length, σ_l and σ_d the standard deviations of length and diameter in the true samples, and r the correlation between length and diameter, and, taking typical cases, they showed that D_l-D_t is of practical significance.

Unless the sampling method is carefully designed to select fibres in either of these precise ways, the amount of length bias is indeterminate and may vary. As Stanbury and Daniels remarked, such methods are best avoided. They recommended that in taking a sample for fineness measurement, the length-biassed type should be used. The reason they gave for this was that "the number of fibres which can be packed together into a given cross-section, and the manner in which the fibres lie together, will depend on the distribution of fibre diameters over the cross-section of the top, sliver or yarn, and this is just what is given by the length-biassed sample. Thus in its relation to spinning power the length-biassed sample is of greater value than the true sample".

This sampling technique was adopted before the war by the International Wool Textile Organisation (Franz, 1937) as part of their agreed standard method for wool-fineness measurement. It is, of course, only applicable to those cases where the wool fibres are arranged in parallel in a riband such as a top, or yarn. The extraction of a sample from an unscoured fleece is a different problem. Because of the great variation in fibre quality from place to place in the fleece, care has to be taken to divide the fleece systematically into suitable sub-samples or zones, and from each of these a small sheaf of fibres is withdrawn to make up a composite sample. The method was described by Wildman and Daniels (1937).

The method of measuring the fineness of a satisfactory sample has given rise to a great deal of experiment. Barker and King (1926) measured weight per unit fibre length. From several places along a sliver or top, small samples were taken between finger and thumb and separated laterally from the rest. A bundle of fibres of suitable size was then prepared, the fibres being laid straight and parallel, and a cutting instrument was used to cut a fixed length from the centre of the bundle. Five hundred of these measured lengths were then counted out, cleaned, and their dry weight found. The method in its essential details has been used by other workers, of whom only Winson (1930, 31) and Barker and Winson (1931) need be mentioned; Stanbury and Daniels (1937) have pointed out its chief defects. It is slow and laborious, the straightening of every fibre in a bundle prior to cutting is very difficult as wool fibres, with their great tendency to return to their previous form when any tensioning force is removed, are extremely intractable; there is a consequent uncertainty in the value of the measured length for weighing. There is also a definite disadvantage in that only a mean value of fineness is given and no measure of the variability of fineness. Apart from

the practical value of a knowledge of this variability it is also desirable because it enables the magnitude of the sampling errors of the mean to be stated in statistical terms. Furthermore, it will be seen that the finger-and-thumb method of extracting a sample gives an indeterminate amount of length bias, and also, if the method is applied to wool containing any proportion of medullated and, therefore, less dense fibres, the average thickness of the fibres is under-estimated (Winson, 1930).

During the last fifteen years the chief methods developed, and almost universally adopted for fineness measurement, employ the projection microscope to form an image of the fibres with a magnification of 500. Such methods have been adopted by the U.S. Government (A.S.T.M., 1944), by the International Wool Textile Organisation (Franz, 1937) and by the Wool Industries Research Association (e.g. Stanbury and Daniels, 1937).

There are two possible methods of measurement : either to measure the profiles of projected images of fibres or to measure the projected images of cross-sections. The A.S.T.M. (1944) allow both methods, whereas the I.W.T.O. (Franz, 1937) have adopted the first method only. A closer inspection of the standard American profile method reveals that the method of selecting the sample is also one in which there is a danger of the amount of length bias introduced being indeterminate. The section method has been criticized on the grounds that there is a danger of the fibre section being distorted during the cutting operation.

Some of the experiments which led up to the eventual formulation of the standard I.W.T.O. method were described by Stanbury and Daniels (1937), and some of their results have already been quoted. Franz (1937) described the details of the technique agreed upon, and from that time until the outbreak of war, refinements were continuously under review in various European laboratories which were co-operating with one another through the I.W.T.O. Henning (1940) visited several laboratories under the auspices of this organisation and examined further important details of technique which were being investigated in order to improve the standard method and to lead to closer agreement in fineness measurement in different laboratories.

Other methods of fineness measurement have been put forward by Ewles (1928) and Cassie (1942). Ewles used the diffraction pattern produced by a sample of fibres, but this has not attracted very much attention, and no developments along these lines can be recorded. The method used by Cassie depends on a theorem of hydrodynamics put forward by Greenhill (1881) and recently adapted by Kozeny (1931) to examine the problem of flow through soils, and by Carman (1937 a–39) to measure the surface area of small particles in gold refining. It has been suggested as a means of measuring the surface area or diameter of textile fibres (Wiggins, Campbell and Maas, 1939 ; Fowler and Hertel, 1940 ; Sullivan and Hertel, 1940 ; Sullivan, 1941), and was used as such by Cassie to measure the diameter of wool fibres. A sample of these was compressed into a tightly packed porous plug of cylindrical form with fibres parallel to the axis of the plug, and the rate of flow of air through the plug, under a known pressure difference, was measured. This rate of flow is a measure of fibre diameter, the formula given by Cassie being an adaptation of Carman's formula to the case of wool fibres. Cassie's investigations showed remarkably close agreement

between the results of this method of diameter measurement and the microscope measurement of diameter for many types of wool, in fact the differences in the results given by the two methods were, in the great majority of cases, less than the sampling and other errors of the experiments; they were, statistically, not significant.

(iv) *Fibre length and wool quality*

Next to fineness, length of fibre and uniformity of length are probably the most important wool characters from the point of view of the influence that fibre properties have on performance during processing. The reason why is discussed at length later in this *Report*, where, because of the intrinsic interest that the subject has, and its practical importance, a full section is devoted to it. The wool sorter, therefore, takes these attributes into account, and, to the best of his ability, selects fibres that are uniform in length for his matchings. Although processing, for example carding, cannot alter the diameter of fibres, it can, and does, alter their length; any operation in which fibres are disentangled or straightened by drawing through pins, or other wool fibres, tends to break them. Thus the sorters' attempts to secure uniformity of fibre length are, to some extent, defeated by the subsequent processing, and, on the other hand, one of the major concerns of combers and spinners is to ensure that this breakage is as small as possible.

An accurate method of length measurement, which would be fairly readily adaptable for routine measurements, would therefore be of great value in investigating the causes of fibre breakage. It would also be most useful for comparing the severity of damage at various points during processing, and for examining the degree of improvement brought about by allegedly better mechanisms.

(v) *Fibre-length measurement*

In spite of the apparent simplicity of the problem of measuring the length of fibres, there are difficulties that are not lightly overcome and accurate results are not easily obtained; this is proved by the fact that no method has yet proved universally acceptable. The reasons for this are, first, that wool has a natural crimp, or curl, and, secondly, it has a great power of elastic recovery from any attempt to straighten it. The consequence is that it is very easily damaged when a sample is removed from bulk, and in the attempt either to comb the fibres in a sample into a straight form for measurement, or to withdraw fibres from a sample for individual measurement, fibre breakage, which vitiates the results, occurs unless minute precautions are observed.

A review of the developments which have taken place in length measurement has been very concisely made by Daniels (1942). He pointed out that the work has followed two different courses. The first, of which there are several variations, consists in principle of clamping a sliver at right angles to its length and removing all fibres which are not gripped, first by drawing the majority gently away by hand, and then removing the remainder with a fine toothed comb. The tuft is then cut off at its bases at the edge of the clamp and weighed. The weight of two tufts divided by the weight per unit length of sliver gives the mean fibre length.

This method was apparently first dscribed by Müller (1894), and has been discussed by Foster, Gregory and Womersley (1933) in connection with the

measurement of cotton fibres. Frenzel (1922), Müller (1923) and Huberty (1938), among others, have shown how the distribution of fibre length in a tuft can be obtained by cutting it into sections of unit length and weighing these separately. More recently Hertel (1940) has adopted it for cotton, and has developed an instrument which measures the tuft thickness along its length, and so the fibre-length distribution in the tuft, by photo-electric means.

From what has already been said it will be recognized that it is practically impossible to prepare these combed tufts without an indeterminate amount of fibre breakage whilst combing ; this is particularly so when the fibres are in a very entangled state, as in card sliver, for which the method is unsuitable. It has other disadvantages, for example, due to the correlation between fibre diameter and fibre length already mentioned (Stanbury and Daniels, 1937), the weight contributed by a long fibre to a tuft is more than proportional to its length, and fibre length is over-estimated. This is perhaps not a very weighty criticism ; another more important one is that, as the result depends on the measured weights per unit length of sliver and sample, it is very sensitive to any lack of straightness of the fibres in either, and gives a measure of the curly or crimped length rather than the straight length.

The other line of approach is that in which the proportion of fibres of different lengths in a sample is found by the direct method of withdrawing them one by one and measuring them. The results depend on the type of sample extracted, and, for any real value to attach to them, this must be either of the true or the completely length-biassed type. Wilkinson (1928) and Townend (1938) both described methods of sampling for the extraction of a length-biassed sample, Wilkinson using a loop of thread, which was inserted into a fibre mass by means of a needle, and drawn tight to encircle a section of them, and Townend using a pair of specially designed surgical forceps for the same purpose. With both methods the danger is that some fibres will be broken when the sample is withdrawn, and others when the tuft is combed to remove loosely adhering fibres.

Daniels (1942) described a new method of length measurement, probably the most accurate technique yet devised, which yields a true sample and obviates the breakage of fibres. Briefly it consists of opening the sliver out into a ribbon-like form, cutting it at right angles to its length, and removing all the fibres from the cut end. The fibres now projecting in the fringe which remains constitute a true sample, and those which terminate in the first millimetre at the tip of the fringe are removed for measurement. Figure 2 shows this being done whilst the wool is held between a glass plate and a velvet padded board. The glass plate, whose weight is adjustable, controls the fibres and ensures that the arrangement of fibres in the web being measured is not disturbed during the operation. With a suitable controlling pressure, provided by the glass plate, and care in withdrawing the fibres, they are not broken. The straight edge of the glass plate affords a means, in the first place, of squaring up the cut end, and secondly, of exposing 1 mm. of fringe for fibre extraction.

Both Townend (1938) and Daniels (1942) give examples of the application of their methods to problems of practical interest to the worsted spinner, and it may be noted that the results of measurements from either type of sample can easily be converted to those of the other type by a simple calculation.

Figure 2. The measurement of fibre length (Daniels, 1942).
(*Reproduced, by kind permission, from the 'Journal of the Textile Institute', Manchester.*)

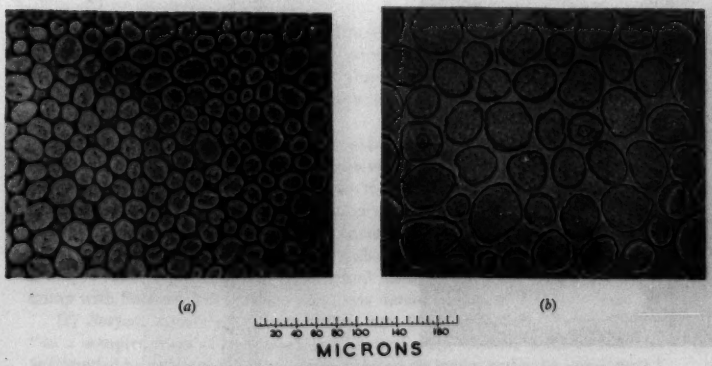


Figure 3. Wool fibre cross-sections, $\times 200$.
(*a*) 70s Australian merino ; (*b*) 46s New Zealand cross-bred (Wildman, 1940).
(*Reproduced by courtesy of the Wool Industries Research Association.*)

To face page 184

Figure 4. Crimp in border Leicester wool (Nicholls, 1932).

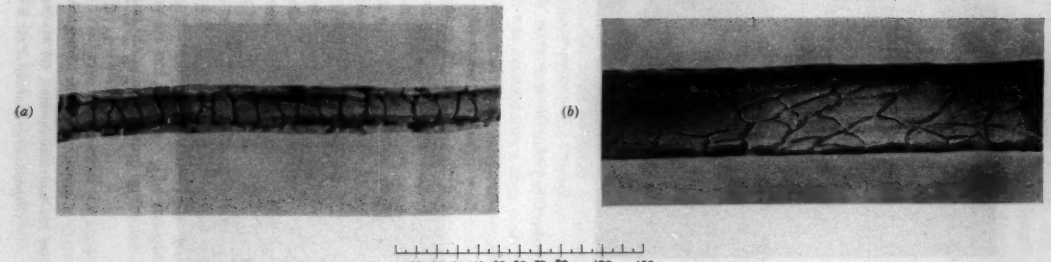


Figure 5. Wool fibre : scale structure ×400.
Impressions formed in resin (Wildman, 1940). (*a*) 80s quality merino ; (*b*) Leicester wool.

(Figures 4 and 5 are reproduced by courtesy of the Wool Industries Research Association.)

(vi) *Fibre shape*

(a) *Cross-section.* Wool fibres are in general elliptical in cross-section (see figure 3), and the "contour ratio" of a fibre at a point along its length is defined as the ratio of the lengths of the major and minor axes of the ellipse at that point. The ratio varies from fibre to fibre, and indeed along each fibre—another example of the infinite variety of wool-fibre properties. Winson (1930) found mean values of contour ratio of from 1·16 to 1·20 in a series of samples ranging from the coarsest to the finest types of wool. Barker (1931) records contour ratios, obtained by Barker and Burgess, ranging from 1·05 to 1·37, and, according to the opinions which spinners gave them of the performance of these wools in practice, they concluded that a low value of contour ratio was a contributory cause of good spinning quality. Barker attempted to explain this by the ability of circular fibres to pack together more compactly than elliptical ones in a strand of yarn, a property which, he said, would enable them to twist more easily and more uniformly during spinning, but it cannot be said that any work has been done since to prove or disprove this contention.

(b) *Crimp.* The longitudinal shape of a wool fibre is characteristic of the material, and one has only to withdraw a fibre from any woollen material to see its crimped form, the crimps being a series of undulations of approximately sinusoidal form (see figure 4). Interesting facts and theories relating to crimp formation have also been summarized by Barker (1931), and it is interesting to mention the observation that crimp formation appears to be a function of time, irrespective of the length of fibre produced in that time, that is to say, adjacent wool follicles on a sheep produce the same number of crimps regardless of the length of fibre they extrude. At the same time as the crimp is being formed the elliptical fibre rotates round its axis as it is being extruded from the follicle; this rotation of the fibre is in phase with the longitudinal crimp wave, i.e. the major axis of the ellipse turns through 360° as one passes from crest to crest of the crimp wave. The explanation, and further elucidation, of these interesting facts, and others not mentioned here, seem to lie rather in the realms of biology than of physics, but the physicist might play his part in explaining the wave formations in terms of the forces brought to bear on the semi-plastic fibre in process of formation.

For our immediate purpose it is not possible to point to any systematic and convincing results on the effect of the degree of crimp on spinning quality, although it is popularly supposed among wool men that it is a desirable feature from this point of view. It is known that the number of crimps per unit length in different qualities of wool is, in general, highly correlated with fibre fineness (Duerden, 1929), but whether the insistence on a well-crimped wool as an index of quality is due to some inherent advantage of crimp itself, or merely to an association of crimp with fineness, has not been effectively demonstrated.

(c) *Surface structure.* A wool fibre has been described (Barker, 1931) "as a complex mass of spindle-shaped cortical cells, with or without medulla, surrounded by a thin membranous layer: the whole held together by imbricated * outer scales".

It is this scaly outer protective sheath (see figure 5) which confers on

* Overlapping like the tiles on a roof.

wool some of its most peculiar, and as yet unexplained, properties. The most important of these are its peculiar frictional properties.

When a wool fibre is rubbed lengthwise between finger and thumb it travels in the direction of its root end as if these scales behaved like a ratchet with teeth pointing in the direction of the fibre tip. The fact that the scales do point in this direction has been responsible for the theory that it is these imbrications which are responsible in a mechanical way for wool fibres becoming matted or felted, and so causing wool fabrics to shrink.

The coefficient of friction is different according to the direction of rubbing, and the fibre is said to possess a "Directional Frictional Effect" (D.F.E.) (Martin, 1944). Speakman and Stott (1931) made use of this fact to define "scaliness", S, by the equation $S = \frac{\tan\theta_2 - \tan\theta_1}{\tan\theta_1}$, where θ_1 and θ_2 were the angles of friction in the two directions when fifty parallel fibres, with the scales all pointing in one direction, and mounted in the form of a miniature violin bow, were allowed to slide on a standard pile fabric. The defects of this method were recently summarized by Speakman, Chamberlain and Menkart (1945). θ_1 varies from one kind of wool to another, the construction of the small violin bows is very laborious, and the apparatus measures static friction whereas dynamic friction would seem to be the operative property when fibres are moving over one another during the processes of manufacture. These three workers have therefore devised a new apparatus, which they style a lepidometer, to measure the frictional properties of individual fibres. In this apparatus "the fibre, with its root end downwards (scales pointing upwards), is placed between two surfaces, which are caused to reciprocate under standard pressure. Fibre travel proceeds until the tension developed in a measuring device, to which the tip end of the fibre is attached, is sufficient to arrest its motion", and in this case the maximum tension developed is taken as a measure of "scaliness".

It is clear that Speakman and his co-workers regard scaliness, i.e. number of scales per unit length and their prominence, as being largely responsible for the frictional properties of wool fibres in a purely mechanical fashion. Martin (1944), however, rejects this, which he calls the ratchet hypothesis, on various experimental grounds. For example he finds that wool fibres exhibit a D.F.E. when rubbed on a hard surface such as polished glass, also that certain keratinous materials, such as finger nails and polished porcupine quills, have no observable imbrications but a definite D.F.E. On the other hand wool fibres treated in certain ways which largely destroy the D.F.E., e.g. by treatment with chlorine or proteolytic enzymes—processes which confer non-felting properties on the fibres—leave the scale structure to all appearances unimpaired. On the basis of his discovery of the piezo-electric properties of all substances which show a D.F.E. he attributes this to a polar arrangement of the molecules in the scales.

The correctness of these rival views is largely of immediate importance in providing a proper explanation of the behaviour of wool, chemically treated or otherwise, under conditions which would normally cause it to felt, and this digression therefore indicates another wide and profitable field for physical research. It might also be noted in passing that the lustre of fibres is also most probably influenced by the scale structure.

It is certain that, whatever may be the mechanism, the scales play an important part in the production of friction of fibre on fibre, or fibre on metal (Barker and Marsh, 1928) and, therefore, they must have an important bearing on many of the operations of manufacture, although scarcely any experimental work has been carried out on this subject. During carding, for example, the fibres are teased out and separated from one another, the operation relying on the frictional forces that enable the wool fibres to be gripped effectively by the fine metal teeth which tease the wool. During the combing or attenuation of slivers into yarn similar considerations must also be involved ; the fibres are drawn through fine metal-toothed combs during gilling and, in the later operations of drawing, during the attenuation of the sliver, fibres are made to slide over one another ; therefore it is not unlikely that their frictional properties may be found to be of some importance in determining their spinning power. In addition, the strength of a yarn is to some extent governed by the mutual clinging or frictional properties of the fibres, and for this reason, also, the ability of a wool to produce a fine yarn that will withstand the stresses of manufacture, i.e. spinning power, must depend on these factors. This is borne out by work at the Wool Industries Research Association, where it has been found that wool fibres treated with chlorine gas before spinning, a treatment which increases inter-fibre friction (Martin, 1944), have a much enhanced spinning power.

It is therefore possible that work of this kind on the frictional properties of fibres may, in the future, have a part to play in the further investigation of the fundamental factors responsible for the spinning quality of wool fibres.

(vii) *Tensile properties of wool*

A great deal of work has been done on the tensile properties of wool, the majority of it having been designed to bring to light the nature of the molecular arrangement within the fibre, and an attempt to describe it in anything like adequate detail would justify a complete report to itself. Shorter (1924) put forward the theory that wool is a hygroscopic colloid, and in its elastic behaviour it exhibits colloidal properties. The full investigation of these properties has to take into account the behaviour of the material under different rates of loading, with different contents of absorbed water and other liquids of higher molecular weight, at different temperatures, and after chemical treatments, which modify the strength and configuration of the molecular structure.

The x-ray work on the structure of wool has shown the molecule to consist of polypeptide chains—molecular chains of amino acids—connected by cross-linkages of various types, of which cystine is the most important. In the unstretched form of the wool fibre the molecular chains are folded or crimped ("α keratin"), but, on stretching, these folds are removed ; completely stretched chains ("β keratin") appear after about 20% extension and thereafter the amount of β keratin increases.

Barker (1931) has given a review of this work, and for later developments the contributions of Astbury, Speakman and their co-workers are available. It is difficult to over-emphasize its importance, since a detailed knowledge of the molecular structure, physical constitution and tensile behaviour of wool is of the utmost value in the understanding of practically every industrial operation to which it is

subjected, whether it be of a chemical or of a physical character. During the operations prior to spinning, in which the physical manipulation of fibres plays such a predominating part, the relation between their elastic properties and their moisture content is of first importance. Increasing moisture content causes a decrease in tensile strength, an increase in the "extension at break", a decrease in Young's modulus and a decrease in the modulus of rigidity. Hardy (1918) found that the mean strength of certain fibres was reduced by about 25% by an increase in the atmospheric humidity from 40% to 80%, and, for the same change, the extension at break was increased from 32% to about 40%. Speakman (1930) and Woods (1940) have examined Young's modulus for small extensions of wool fibres, to which, they say, Hooke's law applies, and agree that it decreases by 75% when a wool, originally dry, is brought into moisture equilibrium with a saturated atmosphere. Speakman found Young's modulus to be $4{\cdot}76 \times 10^{10}$ dynes/cm^2 for Cotswold wool fibres in a state of absolute dryness and $1{\cdot}81 \times 10^{10}$ dynes/cm^2 for fibres in a state of complete saturation. On another occasion Speakman (1929) discovered the even more marked effect of absorbed water on the rigidity modulus of Cotswold wool. In equilibrium with perfectly dry air, the rigidity was found to be $1{\cdot}76 \times 10^{10}$ dynes/cm^2, and it was reduced to a fifteenth of this value by a change from absolute dryness to complete saturation.

The full consequences of these observations on the carding, combing, drawing and spinning operations are by no means fully understood in a scientific sense, although many of the practices followed, and many of the beliefs held in the industry, seem to be capable of explanation by them in general terms. For example there is a popular belief, quite generally known, that the North of England has owned its supremacy in spinning to the dampness of the atmosphere in these parts. In the same way, within the industry itself, it has come to be the general practice to have tops "conditioned" to a regain* of 18 to 19%—a condition which corresponds to equilibrium with an atmosphere of about 83% R.H.—before the operations of drawing and spinning are undertaken. Shorter (1923) refers to empirical rules which have been established that a working atmosphere of 72% R.H. represents the optimum condition for the operations of drawing, whilst a lower value of 55% R.H. represents the optimum condition for the final spinning operation. These rules are followed with complete satisfaction by some spinners, who have the relative humidity in their mills under control by artificial means, and the author has information, acquired by tests carried out under controlled conditions, and evidence of the beneficial effects of drawing at certain optimum humidities on the quality of the yarn spun. It is not surprising that this should be so when it is remembered that the success of the drafting operation depends on the uniform behaviour of the fibres as they are accelerated between the slowly revolving feed rollers and the quickly revolving delivery rollers (figure 1); they must emerge regularly and smoothly, without any tendency to contract or curl unduly when they leave the drafting region between the rollers, where they are subjected to a certain amount of tension. The higher the moisture content, the more flexible the fibres are, the more readily

* Moisture expressed as a percentage of the dry weight of wool. See *Report* by A. B. D. Cassie.

they will adopt a straight form under tension, and, therefore, the more easily they will draw one over the other. When they leave the delivery rollers, being comparatively pliable, there will be less tendency for them to return to a curled or crimped form and cause disturbing rearrangements among the fibres ; the forces tending to hold them straight in a parallel and orderly arrangement will meet with less opposition from fibres with a lower Young's modulus. They will therefore be retained in the sliver in a straighter, more parallel, and, therefore, more suitable form for the next drawing operation. Shorter (1923), on the other hand, puts forward arguments against high humidities. He says "high humidity acts deleteriously by increasing the extensibility of the fibres. During drafting the longer fibres are momentarily subjected to considerable tension, owing to their front ends being seized by the rapidly moving delivery rollers, while their back ends are held by the friction of adjacent slowly moving fibres. The shorter fibres (for reasons which will appear later) will not be subjected to so great a tension. Moreover, the tension will not be uniform—it will be greatest at the front end. The drafting process will therefore strain the fibres in an irregular manner. When the fibres emerge from the delivery rollers, this excess tension is removed and the fibres tend to return to the unstrained state. When the fibres have a low moisture content, the excess tension causes little extension, so that it does little harm. When the fibres have a high moisture content, however, considerable extensions are produced, and the subsequent irregular contraction of the fibres on emergence from the delivery rollers will cause a bad spin and bad yarn".

These two conflicting requirements no doubt explain the optimum working conditions found, and, in addition, the decrease of rigidity with increasing moisture content might possibly facilitate the easy insertion of a high degree of twist into the yarn, without the production of excessive disruptive forces.

(viii) *The electrical properties of wool*

Whilst considering the efficiency with which drawing and spinning can be effected, we are afforded a convenient opportunity of discussing the electrical properties of wool, as they have a considerable bearing on the subject, and are also very markedly affected by moisture content. Any association of fibres, in web, sliver or yarn form, can be almost unmanageable if the fibres become electrostatically charged, and the friction of fibre on fibre, or fibre on metal, during any of the operations described has that effect. If the fibres are poor conductors, the wool is generally left with a positive charge, the fibres repel one another, they refuse to lie in neat parallel arrays, cause a hairy beard on the sliver or yarn, which may even be disrupted by the electrostatic forces between them, and, at the slightest opportunity, they cling to revolving rollers and break down the continuity of the process. At low values of moisture content, wool is a sufficiently poor conductor to display all these troubles, but they gradually disappear as the relative humidity of the surroundings and the moisture content of the wool increases. Earp and Marsh (1933) have made careful measurements of the electrical conductivity of wool and have shown that, as the moisture content increases, conductivity rapidly rises. For every increase in moisture content of 2% there is, for example, a thousand-fold increase in conductivity. Another

interesting feature of their work is the demonstration of the existence of a hysteresis effect for electrical conductivity during absorption and desorption of water by wool.*

It is therefore clear on these grounds also that a sufficiently high value of atmospheric humidity is a necessity for the successful spinning of a wool.

Sufficient has now been said to show that the investigation of the spinning of wool is a complex matter which is influenced by many variables. Indications have already been given of two distinct methods of approach to the problem, although they have not been plainly stated. The first is one in which the workers have visualized an explanation of quality in statistical terms. The ultimate goal of this type of approach would be a regression equation that would enable wool quality to be predicted from those fibre properties found experimentally to be partially correlated with it. The second method, of which examples have been given, as, for example, the attempt to explain the efficiency of drafting in terms of the elastic behaviour of fibres, would attempt to formulate an *a priori* mechanical theory of yarn formation, which would explain the ultimate yarn properties by which the fibre quality is determined. This type of work has been commenced and forms the subject of the next section.

§4. FIBRE ARRANGEMENTS AND THE BEHAVIOUR OF FIBRES DURING PROCESSING

As has already been remarked, the spinning quality of a wool is expressed by reference to the finest yarn which can be efficiently manufactured from it, and this, in turn, is governed by the ability of that yarn to withstand the stresses to which it is subjected whilst being twisted and wound on to the revolving spool prepared for its reception. The strength of a yarn, however, is not only determined by the properties of the individual fibres, such as strength or mutual clinging power, but depends to a great extent on the uniformity of their distribution in the yarn.

The number of fibres per cross-section of a yarn is not constant, even the best machine-made yarn may vary from 0·5 to 1·5 of its mean diameter throughout the length of a few yards. As the yarn is formed twist is inserted, but, because of the variation of the fibre number per cross-section, the number of turns of twist per unit length varies, and a yarn will be found to consist of a succession of softly twisted thick places and tightly twisted thin places with varying degrees of thickness and twist, and, therefore, also of varying strengths. Clearly, as the strength of a length of yarn must be governed by that of its weakest section, the more irregular its thickness the greater will be the probability of a weak place occurring in a given length. Thus the efficiency of spinning, i.e. the spinning quality, will depend on the fibre properties of a wool in so far as these affect the irregularity of thickness of the yarns which can be made from it, and a study of the causes of yarn irregularity, and the factors (fibre properties or otherwise) which influence it, are of great importance for a full understanding of the performance of a particular type of wool.

* More recently Baxter (1943) has examined the temperature coefficient of resistance of wool at constant regain. He deduces from these observations that the conductivity of wool is not due to that of water condensed in the intermicellar channels since the activation energy of a moist fibre is totally different from that of water. It is suggested that absorbed water molecules can be taken as impurity centres, causing the wool to become an electronic semi-conductor.

The causes of yarn irregularity are also of importance to a spinner for other reasons. In the first place abnormally irregular yarns are rejected, because they produce very obvious and objectionable streaks in certain designs of cloth. Unwanted effects of this kind are very detrimental, and great care is taken in the preparation of the yarns to avoid irregularity as far as possible. In fine-knitted garments, such as underwear, irregular yarn will produce an irregular knitted web, which will quickly break down into holes where the yarn is thin. In this respect and in many others, including the mechanical processing of yarns, the presence of thickness irregularities results in a wasteful employment of material, as the potentialities of the great majority of the wool fibres are not realized to the full, the yarn quality being determined by the quality of its weakest parts. As even the best yarns are very far from being uniform, the matter is of fundamental interest, even from a limited practical standpoint; the investigations are not merely important from the point of view of correcting faulty processing.

This problem of yarn uniformity is not confined to worsted yarns, but is equally important in cotton spinning, or the spinning of any other fibrous material where thick slivers are attenuated by drawing machinery into fine yarns. It seems to have been considered first by Balls (1928), who was engaged on the investigation of cotton properties. He remarked on the fact that the particles involved (fibres) had a length of about two thousand times that of their diameter, and went on to say that "the study of the relative movements of such elongated particles is somewhat peculiar to the textile industries". This is certainly true, and accounts for the fact that there is no mathematical theory to account for their relative movements during drafting, which still remain something of a mystery. In many cases involving research into the properties of materials one can often find analogous problems in other fields, or, at least, fundamental principles which can help in forming the basis of a mathematical theory, but in this case no similar problem seems to have presented itself elsewhere, and the fundamental laws of fibre movement have not yet been formulated.

Balls (1928) investigated the thickness variations in cotton yarns, and slivers from earlier drafting operations, by cutting them into short lengths and studying the graphical records of the variation of weight per unit length. He found that these variations had a wave-like character, and he introduced the term drafting-wave to describe this phenomenon. Whether the hypothesis of wave-formation put forward by Balls is really justified will be discussed as the subsequent research work is described, but this was the working hypothesis on which a good deal of investigation proceeded.

Balls found that each operation of drafting attenuated the waves which had been formed at earlier stages, and at the same time was responsible for the generation of a new set of waves which were superimposed on those already present. Each operation produced waves of approximately the same length, but, as the slivers drafted became thinner, the amplitude of the waves formed increased. The final qualitative picture of weight per unit length variations in yarn, given by Ball's observations, is that of short-wave variations of considerable amplitude (giving rise to visible thick and thin places) produced in the later processes, particularly the last one, and variations of longer wave-length, and smaller amplitude, produced in the earlier stages of attenuation as short-term variations, and drawn

out to longer and longer lengths as drafting proceeds. All these variations are superimposed on one another, and matters are made still more complicated by the fact that during the drawing operations it is customary to draft several slivers simultaneously at each stage ; thus three "ends" of sliver may be put up together,* the three "ends" emerging from the delivery rollers are united, and, given a draft of six, a two-fold reduction will have been effected by the operation. This doubling is done, of course, to minimise irregularity, but it makes the final pattern of yarn variation much more complex than would be the case if the process consisted of the successive reduction of single ends.

Balls (1928) explained the formation of the drafting-waves in the following way. When a sliver is being drafted, the feed rollers slowly feed forward an approximately parallel group of fibres of varying lengths, and the longest of these will be under their direct control until an instant before their leading ends reach the quickly moving delivery rollers. The shorter fibres, however, are for an appreciable time out of direct control by either set of rollers—they can be said to be floating fibres—and the point within the drafting zone at which they are suddenly accelerated depends on the speed changes of the fibres with which they are most closely linked by frictional forces. In order that the original arrangement of fibres should be preserved, and the uniformity of the sliver thereby maintained as far as possible, all fibres should be accelerated in sequence in the order in which they originally occur in the ingoing sliver, that is to say they should continue to move with the surface speed of the feed rollers until they are positively gripped by the delivery rollers. This does not happen; as Balls pointed out, some of the short fibres are attached to the quickly moving fibres, which are being withdrawn by the delivery rollers, and are swept out of the drafting zone with them to form a thick place. This plucking action is possibly cumulative ; as more fibres move with the speed of the delivery rollers, so will the disturbing influence of these rollers be increased, and still more short fibres may be removed. Such behaviour, however, will eventually lead to a deficit of short floating fibres in the drafting zone, so that the thick place will be followed by a thin place, which will continue until the deficit is made good from a new supply brought forward by the feed rollers ; conditions will then be restored for a repetition of the action.

In this way Balls conceived a mechanism which could give rise to pulsations of thickness due to a periodic removal and replenishment of short fibres. It also seems reasonable to visualize the wave-length and amplitude of the resultant waves being dependent on (i) the relative roller speeds, (ii) the relation between roller spacing and fibre length, and (iii) the variability of fibre length, because all these factors affect the relative rates of removal and replenishment ; but no mathematical theory has yet been formulated to explain the exact mode of fibre movement in such circumstances. The way in which the number of fibres in the drafting zone (thickness of sliver) might be expected to affect the amplitude of the thickness waves is not at all obvious, nor is the part to be played by inter-fibre friction, because, although it might be said that increased frictional forces would give the quickly moving fibres a better chance of removing short floating fibres, it can equally well be argued that the slowly

* This operation is known technically as *doubling*, and the number of ends so combined as the number of doublings.

moving fibres, under the direct control of the feed rollers, would have a proportionally better chance of retaining them in their proper position.

Following on this work by Balls, Goodings (1931) carried out some experiments on worsted slivers and yarns, and he agreed that the same general considerations could be said to apply to the drafting of wool fibres in the worsted process. The author of this *Report*, in work carried out at the Wool Industries Research Association (Martindale, 1942), also testified to the correctness of the general views so far put forward as they apply to worsted drawing.

Much work has been done at the British Cotton Industry Research Association on this subject as it applies to the processing of cotton, and, in his first paper bearing on it, Foster (1929) also accepted the theory of floating fibres put forward by Balls, and confirmed his observations that each operation generated a fresh set of waves ; also that their amplitude increased as the drawing proceeded and the slivers drafted were reduced in thickness. Foster brought a considerable amount of further information to light, which will be discussed in a later paragraph, but at this point, having introduced the nature of the problem to be investigated, a convenient opportunity is presented to mention briefly the instrumental methods by which irregularity of slivers and yarns has been examined.

The first and most obvious means which was adopted was to cut the materials into short sections of convenient length and weigh them, this being the method employed by Balls (1928) and Goodings (1931) ; but with the great productive power even of experimental textile machinery, and the great number of experiments which could be done very simply, these methods were much too slow. Furthermore, with such problems of variability, statistical considerations demand a large number of observations in order to achieve any degree of accuracy, and, therefore, instruments have been designed to give records of irregularity in either graphical or tabulated form.

Of these indirect methods used for measuring the variation of thickness of yarn, the following list, which is not exhaustive, gives some idea of the variety of techniques adopted : photo-electric measurement of thickness with the use of manual plotting apparatus for following the magnified deflections of a Lindemann electrometer (Barker and Stanbury, 1928 ; Stanbury, 1931); measurement by means of a thermopile (Whitehead, Steward and Smith, 1929); measurement by diffraction effects (Matthew, 1932); measurement of the mode of vibration of yarns (Richard, 1928); measurement of the change of resistance of a mercury capillary through which the yarn is passed (Viviani, 1933); and several instruments in which the thickness is measured by a light mechanical feeler, whose movement is magnified mechanically or optically (Saxl, 1935 ; Froning, 1936 ; Cavaney, Foster and Gregory, 1936 a ; Foster and Womersley, 1936). Recently Martindale (1945 a) has described a modified cutting method in which the weighing of cut unit lengths is again resorted to ; n yards of yarn are wrapped on a yarn-wrapping reel and a clamp with a one-inch face is used to hold the wrapped hank whilst a one-inch section of it is cut out. The difference between the aggregate weight of the n one-inch lengths and the mean weight per inch length of the complete hank is a measure of the irregularity of the "n-fold" yarns, which is more accurately determined by testing 100 hanks in the same way. If σ_n is the standard deviation of weight per unit length of the n-fold hank then $\sigma_n/\sqrt{n}$ is that of the single yarn.

Some of these methods give a complete graphical representation of yarn-thickness variation, whilst others, which merely give an estimate of the standard deviation of thickness, only provide a measure of the amplitude of the variations ; whilst the former is necessary for a complete analysis of the problem of wave-forms and fibre movement, the second type can give sufficient information in the investigation of many of its aspects.

Saxl (1935) received a beam of light from the yarn–feeler mechanism on a photo-cell, and, by leaving this cell for a fixed interval of time in various positions in the path of the oscillating beam, he was able to build up a frequency distribution of yarn thickness. The instrument described by Cavaney, Foster and Gergory (1936 a) gives a continuous photographic record of thickness ; in a second paper (Cavaney, Foster and Gregory, 1936 b) they discuss the principles of yarn-thickness measurement in general and the operation of their machine in particular. They point out that as yarn contains softly twisted, easily compressible, thick places and highly twisted, and relatively incompressible, thin places, a light pressure is necessary to give a faithful record of thickness. Their object was to obtain records whose dimensions were highly correlated with weight per unit length, and they found that a feeler pressure of 2·5 gm. gave the necessary high degree of correlation. It is necessary, of course, before one can use graphical records of this kind to deduce the behaviour of the fibres, to ensure that the property of the yarn measured is highly correlated with weight per unit length in this way.

The apparatus built by Foster and Womersley (1936) is the most elaborate yarn-irregularity tester yet built. It consists of a battery of six units and tests six yarns simultaneously. Each unit is fitted with a light mechanical feeler mechanism of the type found so useful on the continuous photographic recorder, but in this case they work intermittently, taking snap readings of thickness at points about a yard apart, as yarn is drawn through the instrument. Each engagement of the feeler with the yarn sets the measuring levers associated with it in accordance with the yarn thickness, and the set distance of each of these levers from the zero position is measured by a metal rod which advances from the zero position to the set lever. As this rod advances it rotates two discs coupled to it by friction clutches ; one of these records the distance moved, the other the square of this distance. At the moment of contact with the feeler mechanism a current flows through an associated thyratron relay and two solenoids de-clutch the rotating discs, one of which therefore shows a reading proportional to yarn thickness, the other a reading proportional to the square of the thickness. After re-setting, the cycle of operations is repeated and the discs integrate the series of observations. From the integrated results the standard deviation of yarn thickness is readily calculated. Lack of space prevents a fuller description of this instrument, which incorporates an ingenious mechanical squaring mechanism. This, and the performance of the integrating mechanism, is fully described in the original paper. Another paper by Foster (1936) will also be of interest to anyone who may be concerned with the design of mechanical calculating devices, which can be fitted to testing machines, for the automatic computing of statistical coefficients.

Yarn of course is the product of the final drawing operation, but, as has been

explained already, the earlier drafted products, the slivers, are not tightly twisted—they are soft and equally compressible at all points ; there are slight variations of twist in them, but this has very little effect on compressibility, and the measurement of their thickness (weight per unit length) variation is comparatively simple. It is customary to pass them through a rectangular slot and compress them by a loaded tongue or blade into a rectangular section. If a continuous record is required, a rectangular groove is cut in the periphery of a wheel, and a loaded tongued wheel rides on the compressed sliver in the groove. The position of the blade in the one case, or of the tongued wheel in the other, is magnified by suitable means and made to provide snap readings, or a continuous record, in any convenient manner. Cavaney, Foster and Gregory (1936 a, 36 b) described a photographically recording sliver tester designed on these lines, discussed the details of its performance, and verified the close correlation between the records and records of weight per unit length. Vose (1944) and Plummer (1945) described an American machine designed on similar lines and its application to problems of the irregularity of cotton slivers. Martindale (1942) gave details of an instrument which works on these same principles, taking snap readings of thickness at points 9 in. apart along worsted slivers. He showed that a high degree of correlation existed between the instrumental reading and weight per unit length of these slivers, as distinct from cotton slivers on which all other instruments described have operated. Cavaney and Foster (1937) have described an automatic sliver tester for cotton consisting of a battery of six units, each measuring the thickness of slivers intermittently at points about 10 in. apart along their length. This is a companion instrument to the automatic yarn-irregularity tester, and includes automatic calculating devices working on the same principles as those designed for that machine, but constructed on more robust lines. Again they discuss the performance of the mechanism in detail, and verify its accuracy in measuring the standard deviation of weight per unit length of slivers.

With instruments of this kind the study of the drafting waves produced under a wide variety of experimental conditions was greatly accelerated.

Balls (1928) was the first to undertake the analysis of these thickness variations, but he soon found that the problem was so complex, and the amount of computation involved so great, that he had to design an instrument to perform the periodogram and harmonic analysis mechanically (Balls, 1921). In this he was followed by Foster (1930), who designed his grating periodograph to perform the analysis of photographic thickness records by optical means. With his mechanical aid Balls found the real situation even more difficult than he had anticipated. He had proceeded from the assumption that the variations would be found to consist of random variations superimposed on certain underlying fundamental periodicities, but after his investigations he commented that "periodic variations certainly existed, but in excess. The length of waves was so variable that some could be detected at all lengths between 1 mm. and 50 cm. in the same yarn. Discrimination between these various lengths with respect to their amplitude and their form, and the analysis of that form into harmonics, presented a colossal task". He concluded that "progress in the study of drafting will continue to be slow until the construction of the drafting wave is understood analytically".

Foster (1929) discovered similar confusion when he carried out similar investigations with the grating periodograph. He found that if a length of yarn was examined in successive sections, a number of periodicities were present in each section, but they were not common to all sections, sometimes they cropped up again in other sections and sometimes they did not. All periods found in a section appeared to be harmonics of a fundamental period, the particular harmonics present in a section varied in a haphazard way, and the length of the fundamental period depended on the length of material measured ; it was comparable with, but rather shorter than, the length of the sample analysed. Foster showed that all these results could be accounted for by discarding the previous conception of a regular periodic form of the drafting wave and considering it as a wave motion continually varying in amplitude and phase.

This new conception of the wave motion led to the consideration of work which had been carried out in other fields. The form of variation which is found in drafted products, e.g. figure 6, resembles data which have accumulated in other ways, for example the time series representing variation in sun-spot

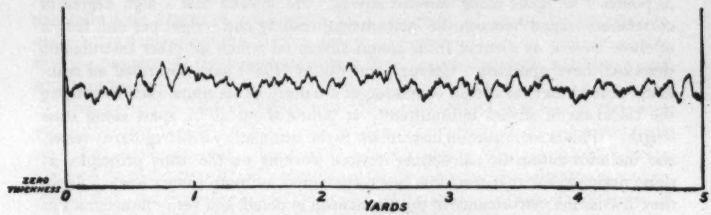


Figure 6. A typical example of thickness variations in worsted slivers, showing the wave-like character of much of the irregularity.

activity (Yule, 1927), meteorological data (Walker, 1931), or economic time series (Davis, 1941), and a good deal of analysis has been done from time to time in attempts to establish the identity and reality of cycles of activity in these fields. Davis gives an excellent historical account of the way in which ideas have crystallized regarding the nature of time series of these kinds, and the analytical methods developed to "disentangle the structural from the erratic in them". It is recorded by Davis that many of the time series, when subjected to ordinary periodogram analysis, displayed exactly the same features as the yarn and sliver thickness records examined by Foster (1929), for example in the case of sun-spot numbers between the years 1750 and 1900 about 35% of the total variation is accounted for by a single period of 11·25 years ; for the first half of this interval, however. this period is absent and periods of 9·25 years and 13·75 years appear; during the second half (1826-1900) 85% of the variation is again accounted for by the 11-year period. It may be noted that all these are harmonics of a fundamental of about 55 years, a length comparable with, but rather shorter than, the length of the interval analysed.

Yule (1927) criticized the tendency to start from the hypothesis that true periodicities are present in such data "masked solely by more or less random superposed fluctuations—fluctuations which do not in any way disturb the steady course of the underlying periodic function or functions". If such were the true state of affairs, periodogram analysis would detect these periodicities, and the time series would in many cases show some sign of regular periodicities if these were of any great importance, whereas in most cases neither the time series nor periodogram analyses display any persistent periodicity. Yule showed that if, on the other hand, the fortuitous disturbances were not superposed, but affected the vibrations themselves, an entirely different result might be expected. He cited as a simple analogy the case of a pendulum "pelted with peas, sometimes from one side and sometimes from the other", and showed that its motion would remain very smooth, but would vary continuously in both amplitude and phase, whilst the resulting graphical record would preserve an approximately sinusoidal appearance, to an extent dependent on the magnitude of the fortuitous disturbances.

Another method of analysis has been developed to analyse data of this kind—the method of auto-correlation (Davis, 1941). If the series consists of readings (say of yarn thickness) $u_1, u_2, \ldots u_{n+s}$ at equal intervals of t, and $y_1, y_2, \ldots y_{n+s}$ are their respective deviates from the mean value, the coefficient of correlation between u's separated by s intervals is given by

$$r_s = \sum_{1}^{n} \frac{y_1 y_{1+s}}{n \sigma_0 \sigma_s},$$

where σ_0 and σ_s are the standard deviations of u_1 to u_n and u_{1+s} to u_{n+s} respectively. If the u's tend to vary periodically with a period λ, the mean length of the disturbed period will tend to approach λ if the $n+s$ observations contain a large number of periods; for values of s equal to 0, λ, 2λ, etc., the value of r_s will have maximum positive values (when $s=0$, $r_s=1$) and if $s=\lambda/2$, $3\lambda/2$, etc., r_s will be a minimum. Plotting r_s against s gives the correlation periodogram or correlogram.

Martindale (1941) described an optical instrument for the automatic plotting of these periodograms, as the labour of computing them for the large amount of data being collected in the investigation of the drafting problem was prohibitive.* In this instrument the serial data are presented in duplicate form in the shape of two templates, A and B (figure 7), with different longitudinal scales. If these scales are in the ratio $a : b$, the templates are erected parallel to one another in vertical planes at distances $k.a$ and $k.b$ respectively from a narrow vertical slit P; the large curve A, which is therefore at the greater distance from the slit, is illuminated by diffuse light. If a "fan" of rays from a vertical element, of width δt, at the point u_t in A reaches the slit by passing through B at u_{t+s}, the amount of light falling on the slit is proportional to $u_t u_{t+s} \delta t$, and the total light from the whole curve, comprising n observations,

$$I_s = C\,\delta t \,.\, \sum_{1}^{n} u_t u_{t+s}.$$

* A symposium on *Autocorrelation in Time Series* is being held by the Research Section of the Royal Statistical Society in January 1946. At this symposium a review is being made of machines for computing correlograms.

If y_l is the deviation of u_l from its mean value $\bar{u}_0$, determined over the observations u_1 to u_n, and if y_{l+s} is the deviation of u_{l+s} from its mean value $\bar{u}_s$, determined over the range of observations u_{1+s} to u_{n+s}, then

$$I_s = C\,\delta t \,.\, \sum_{l=1}^{n} (\bar{u}_0 \bar{u}_s + y_l y_{l+s})$$
$$= n\,C\,\delta t \,.\, (\bar{u}_0 \bar{u}_s + r_s \sigma_0 \sigma_s)$$

If s is small compared with n, $I_s = C_1 r_s + C_2$.

The variation of I_s as the curve B is traversed parallel to A is thus proportional to r_s, and this amount of light passing through the slit is focused on a photo-cell. The amplified output of the photo-cell is then plotted, as B is traversed, to construct the periodogram.

With this correlation periodograph, the periodogram for sun-spot numbers was obtained as far as the first maxima ($s = \pm\lambda$) on both sides of the central

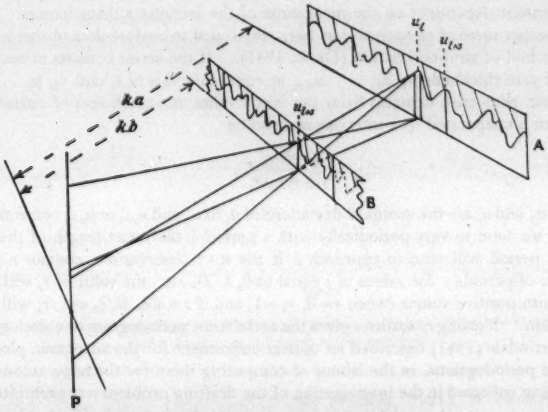


Figure 7. Correlation periodograph (Martindale, 1941).

(Reproduced, by kind permission, from the 'Shirley Institute Memoirs'.)

maximum ($s = 0$), and it gave the value of the period as 10·7 years compared with Yule's computed value of 10·6 years. The correlation periodogram for the pressure measurements at Port Darwin, examined by Sir Gilbert Walker (1931), gave the value of 11·3 quarters compared with Walker's value of "about 11¼ quarters".

Foster's conclusion (1929) that the drafting wave varied in both amplitude and phase seemed to put it in the same category as Yule's analysis placed the variation in sun-spot number, that of a wave disturbed by random disturbances. Foster and Martindale (1941) consequently examined typical cases of drafted cotton slivers and found that the correlation periodograms obtained were consistent with this hypothesis. They therefore suggested that the drafting wave was the product of a system which tended to oscillate with a natural frequency

of its own, but that it was continually being disturbed by random, erratic shocks. The tendency to oscillate with a certain natural period was explained by the tendency towards a regular periodic removal and replenishment of short floating fibres in the drafting zone by the mechanism already outlined. The erratic disturbances, it was suggested, might be provided by inequalities in the material presented for drafting, such as inequalities in fibre number, fibre distribution, parallelisation of fibres, etc. Altogether periodograms were constructed for 36 drafted slivers, prepared at different drafts and different roller spacings from normally twisted slivers of three types of cotton. In all, except some of those prepared with low drafts and close roller spacings, the first minima and first maxima of the periodograms ($s = \pm\lambda/2$ and $s = \pm\lambda$) were clearly marked, but beyond $s = \pm\lambda$ the correlation became very low. The conclusion was therefore drawn that the disturbances were so great that the original oscillation was obliterated after about one period. It therefore becomes an open question whether such a phenomenon can really be dignified by the term drafting wave under normal

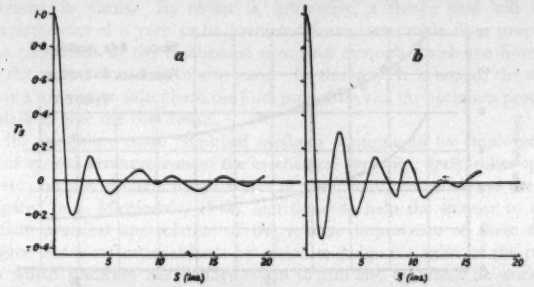


Figure 8. Correlation periodograms of cotton sliver variations (Foster and Martindale, 1941). (*a*) Normal twist in drafted sliver ; (*b*) high twist in drafted sliver.
(*Reproduced, by kind permission, from the 'Shirley Institute Memoirs'.*)

conditions, but it was shown that the tendency to oscillate with a natural period did exist, because under more favourable conditions, wider roller spacings, higher drafts, and in particular higher degrees of twist in the material presented for drafting, the wave motion became more pronounced and was not damped out so rapidly. Evidently the higher twist either increases the force reponsible for the fixed period, or decreases the effect of the disturbing forces ; the result is that the phase and amplitude changes are smaller and the drafted sliver has a more regular periodic variation in thickness.

Foster and Martindale tabulated the effect of these variables, draft, roller spacing and twist per unit length, on the length of the wave for different types of cotton, but, so far as the writer is aware, the results have not yet been used to advance any mathematical theory of drafting ; nor has any comparable work been done with other fibres.

More recently, other evidence has been brought to light confirming the reality of some kind of periodicity in drafted slivers (Spencer-Smith and Todd,

1941 ; Martindale, 1945 a). Spencer-Smith and Todd were interested in flax slivers and, assuming a Poisson distribution of fibre ends, calculated the irregularity of thickness of a sliver in which the fibres were distributed at random. They obtained the formula $C = 0{\cdot}905\sqrt{AY}$, where C was the coefficient of variation of thickness, A the mean area of fibre cross-section in $10^{-5} \times \text{cm}^2$ and Y the weight of the sliver expressed as yards per ounce. All slivers measured had a value of the coefficient significantly greater than this and, furthermore, Spencer-Smith and Todd found that the "correlogram" (correlation periodogram) from actual slivers was significantly different from that which would be expected from a sliver of randomly arranged fibres. In such a sliver they show that this correlogram would merely take the form of a second integral of the fibre-length distribution due to the occurrence of the same fibre in adjacent sections of sliver. Both these results confirmed that some kind of periodicity was superposed on the random structure of the sliver.

Martindale (1945 a) considered the formation of worsted yarns, and pointed out that for a yarn to be perfectly uniform the necessary condition is that the fibres

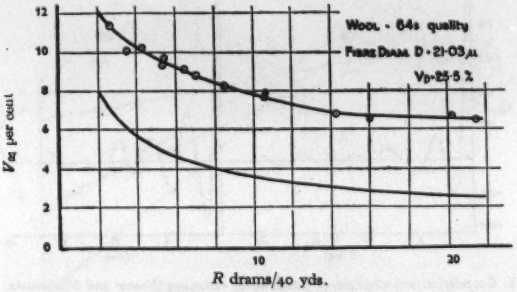


Figure 9. (Martindale, 1945 a.)

(*Reproduced, by kind permission, from the 'Journal of the Textile Institute', Manchester.*)

should be arranged in orderly fashion, so that every time a fibre ended another must immediately be introduced to take its place. As the drawing machinery has not the necessary faculty of perception to ensure this, the best that any set of drafting rollers can do is to preserve the order of fibres presented to it, and it was shown, by a consideration of the processes preceding gilling and drawing, that at best this could only be a random order. The most perfect sliver or yarn that can be made, therefore, by present methods, is one in which the fibres are arranged at random. The formula given by Martindale to represent this lower limit of irregularity is, in the case of worsted slivers or yarns, $V_a = 100(1 + 0{\cdot}0004V_D^2)^{\frac{1}{2}}/\sqrt{n}$, where V_a is the coefficient of variation of sliver or yarn c.s.a., V_D the coefficient of variation of fibre diameter and n is the mean number of fibres per c.s.a. (cross-sectional area).

It was shown for a large number of yarns of different thicknesses, made from different types of wool, that this value was always exceeded. The additional

irregularity was accounted for by the imperfect drafting machinery superposing almost periodic arrangements on a random fibre order. Figure 9 shows results obtained by Martindale on normally manufactured slivers of 64's quality wool ; the top curve shows experimental results obtained relating the coefficient of variation of weight per unit length, V_R, to the weight per unit length, R. The bottom curve shows the relation which would exist if all slivers consisted of a random assemblage of parallel fibres. It seems significant that the two curves follow parallel courses, a result which suggests that whereas drafting waves may be responsible for the major part of irregularity in thick slivers, their contribution to irregularity remains constant, and that the greater part of yarn irregularity is not due to this effect at all, but simply to the inability of a small number of fibres in the section of a yarn to provide the necessary amount of overlapping to preserve the uniformity, unless they are arranged in one particular orderly fashion which the machinery has not the necessary faculties to arrange.

This is as far as any published work will allow us to go in a review of the work which has been done to lay bare the fundamentals of fibre movement and the fibre arrangement in yarns. Its object is, ultimately, a theory that will enable the characteristics of a yarn to be predicted from measurable fibre properties, and the particulars of the mechanical processes through which the fibres pass during the attenuation of slivers to yarn. In this way, it is hoped, the spinner will be in a position to select both the fibre properties and the optimum processing particulars to give the best result.

In the meantime more empirical methods continue to be employed ; the effects of various arrangements of the mechanical variables, draft, roller spacing, twist, etc., on the ultimate irregularities of sliver and yarn thickness are being investigated (e.g. Martindale, 1942), and these do help the spinner to a more immediate practical appreciation of the relative importance of these factors. They give him some rational basis by which to judge the value of the rules of thumb which tradition has handed down to him and by which he works, and have already had the beneficial effect of getting some spinners to regard these in a more critical light. One result has already been that some worsted spinners have now gained sufficient confidence to institute mill trials of their own to assess the practical value of some of the results brought to light in these empirical investigations, and that, in a traditional industry, constitutes some progress for stimulating which physics and the advocates of reliable physical methods of enquiry can take some credit.

ACKNOWLEDGMENT

The author's thanks are due to the Council of the Wool Industries Research Association for permission to contribute this *Report*.

REFERENCES

American Society of Testing Materials, 1944. *Standards on Textile Materials*, p.315.
Balls, W. L., 1921. *Proc. Roy. Soc.*, A, 99, 283.
Balls, W. L., 1928. *Studies of Quality in Cotton* (London : Macmillan), ch. 8.
Barker, S. G., 1931. *Wool Quality* (London : H.M. Stationery Office).
Barker, S. G. and King, A. T., 1926. *J. Text. Inst., Manchr.*, 17, T. 68.
Barker, S. G. and Marsh, M. C., 1928. *Wool Ind. Res. Assoc.*, Pub. No. 89.

BARKER, S. G. and STANBURY, G. R., 1928. *J. Text. Inst., Manchr.*, **19**, T. 405.
BARKER, S. G. and WINSON, C. G., 1931. *J. Text. Inst., Manchr.*, **22**, T. 314.
BAXTER, S. 1943. *Trans. Faraday Soc.*, **39**, 207.
VON BERGEN, W., 1939. *Text. Research*, **10**, 63.
CARMAN, P. C., 1937 a. *Trans. Inst. Chem. Engrs.*, **15**, 150.
CARMAN, P. C., 1937 b. *Trans. Inst. Chem. Engrs.*, **16**, 168.
CARMAN, P. C., 1938. *J. Soc. Chem. Ind.*, **57**, 225.
CARMAN, P. C., 1939. *J. Soc. Chem. Ind.*, **58**, 1.
CASSIE, A. B. D., 1942. *J. Text. Inst., Manchr.*, **33**, T. 195.
CAVANEY, B., FOSTER, G. A. R. and GREGORY, J., 1936 a. *Shirley Inst. Mem.*, A, **4**, 305.
CAVANEY, B., FOSTER, G. A. R. and GREGORY, J., 1936 b. *Shirley Inst. Mem.*, A, **4**, 313.
CAVANEY, B. and FOSTER, G. A. R., 1937. *Shirley Inst. Mem.*, A, **4**, 375.
DANIELS, H. E., 1942. *J. Text. Inst., Manchr.*, **33**, T. 137.
DAUBENTON, 1779. *C.R. Acad. Sci., Paris*; see also BARKER (1931).
DAVIS, H. T., 1941. *The Analysis of Economic Time Series* (Bloomington, Indiana : The Principia Press).
DUERDEN, J. E., 1929. *J. Text. Inst., Manchr.*, **20**, T. 93.
EARP, K. E. and MARSH, M. C., 1933. *Trans. Faraday Soc.*, **29**, 173.
EWLES, J., 1928. *J. Text. Sci.*, **2**, 101.
FECHNER, G. T., 1860. *Elemente der Psychophysik* (Leipzig : Breitkopf and Härtel)
FOSTER, G. A. R., 1929. *Shirley Inst. Mem.*, A, **2**, 313.
FOSTER, G. A. R., 1930. *J. Text. Inst., Manchr.*, **21**, T. 18.
FOSTER, G. A. R., 1936. *J. Text. Inst., Manchr.*, **27**, T. 37.
FOSTER, G. A. R., GREGORY, J. and WOMERSLEY, J. R., 1935. *Shirley Inst. Mem.*, A, **4**, 183.
FOSTER, G. A. R. and MARTINDALE, J. G., 1941. *Shirley Inst. Mem.*, A, **5**, 125.
FOSTER, G. A. R. and WOMERSLEY, J. R., 1936. *Shirley Inst. Mem.*, A, **4**, 325.
FOWLER, J. L. and HERTEL, K. L., 1940. *J. Appl. Phys.*, **11**, 496.
FRANZ, E., 1937. *J. Text. Inst., Manchr.*, **28**, P. 368.
FRENZEL, W., 1922. *Leipz. Mschr. Textilindustr*, **37**, 3.
FRONING, A., 1936. *Leipz. Mschr. Textilindustr.*, **51**, 141.
GOODINGS, A. C., 1931. *J. Text. Inst. Manchr.*, **22**, T. 1.
GREENHILL, A. G., 1881. *Proc. Lond. Math. Soc.*, **13**, 43.
HARDY, J. I., 1918. *J. Agric. Res.*, **14**, 285.
HENNING, H. J., 1940. *Z. ges. Textilindustr.* **43** 36.
HERTEL, K. L., 1940. *Text. Research*, **10**, 510.
HUBERTY, A., 1938. *International Wool Textile Organisation, Technical Committee Report.*
KOZENY, I. J., 1931. *Wasserkraft und Wasserwertschaft*, **1**, 67.
MARTIN, A. J. P., 1944. *J. Soc. Dy. Col , Bradford*, **60**, 325.
MARTINDALE, J. G., 1941. *J. Text. Inst., Manchr.*, **32**, T. 71.
MARTINDALE, J. G., 1942. *J. Text. Inst., Manchr.*, **33**, P. 9.
MARTINDALE, J. G., 1945 a. *J. Text. Inst., Manchr.*, **36**, T. 35.
MARTINDALE, J. G., 1945 b. *J. Text. Inst., Manchr.*, **36**, T. 213.
MATTHEW, J. A., 1932. *J. Text. Inst., Manchr.*, **23**, T. 55.
MÜLLER, E., 1894. *Z. Ver. dtsch. Ing.*, **38**, 997.
MÜLLER, E., 1923. *Leipz. Mschr. Textilindustr.*, **38**, 346.
NICHOLS, J. E., 1932. *Study of Empire Wool Production* (Leeds : Wool Industries Research Association).
PLUMMER, C. H., 1945. *Text. Research*, **15**, 10.
RICHARD, J., 1928. *J. Phys. Radium*, **6**, 775.
SAXL, I. J., 1935. *J. Text. Inst., Manchr.*, **26**, T. 77.
SHORTER, S. A., 1923. *J. Soc. Dyers and Col.*, **39**, 270.
SHORTER, S. A., 1924. *J. Text. Inst., Manchr.*, **15**, T. 207.
SPEAKMAN, J. B., 1929. *Trans. Faraday Soc.*, **22**, 178.
SPEAKMAN, J. B., 1930. *J. Soc. Chem. Ind.*, **49**, T. 209.
SPEAKMAN, J. B., CHAMBERLAIN, N. H. and MENKART, J., 1945. *J. Text. Inst., Manchr.*, **36**, T. 91.
SPEAKMAN, J. B. and STOTT, E., 1931. *J. Text. Inst., Manchr.*, **22**, T. 339.
SPENCER-SMITH, J. L. and TODD, H. A. C., 1941. *J. Roy. Stat. Soc., Suppl.*, **7**, 131.
STANBURY, G. R., 1931. *J. Text. Inst., Manchr.*, **22**, T. 385.

STANBURY, G. R. and DANIELS, H. E., 1937. *J. Text. Inst., Manchr.*, **28**, T. 188.
SULLIVAN, R. R., 1941. *J. Appl. Phys.*, **12**, 503.
SULLIVAN, R. R. and HERTEL, K. L., 1940. *Text. Res.*, **11**, 30.
TOWNEND, S., 1935. *J. Text. Inst., Manchr.*, **26**, T. 130.
TOWNEND, S., 1938. *J. Text. Inst., Manchr.*, **29**, T. 55.
VIVIANI, E., 1933. *Kunstseide*, **15**, 30.
VOSE, R. W., 1944. *Text. Research*, **14**, 76.
WALKER, Sir G., 1931. *Proc. Roy. Soc.*, A, **131**, 518.
WHITEHEAD, W. & J., Ltd., STEWARD, O. M. and SMITH, C., 1929. B.P. 343,215.
WIGGINS, E. J., CAMPBELL, W. B. and MAAS, O., 1939. *Canad. J. Res.*, B, **17**, 318.
WILDMAN, A. B., 1940. *Animal Fibres of Industrial Importance* (Leeds: Wool Industries Research Association).
WILDMAN, A. B. and DANIELS, H. E., 1937. *J. Text. Inst. Manchr.*, **28**, T. 202.
WILKINSON, H., 1928. *J. Text. Sci.*, **2**, 104.
WINSON, C. G., 1930. *J. Text. Inst., Manchr.*, **21**, T. 524.
WINSON, C. G., 1931. *J. Text. Inst., Manchr.*, **22**, T. 533.
WOODS, H. J., 1940. *Leeds Phil. and Lit. Soc.*, **3**, 577.
YULE, G. U., 1927. *Phil. Trans.*, A, **226**, 267.

ELECTRON MICROSCOPY

By L. MARTON,
Division of Electron Optics, Stanford University, California

§1. INTRODUCTION

SINCE the inception of electron microscopy, two accounts of it have appeared in these *Reports*, the first by Starks in 1935 and the second by Klemperer in 1940. The progress made since the last *Report* on this subject, although very considerable, is perhaps less spectacular than the earlier achievements. It is marked by consolidation of the first results, by commercialization of some prototypes of electron microscopes, and by a wide variety of applications in many fields of research.

The present *Report* can be divided into four main parts. The first will deal with developments of the theory of short-focus lenses and the comparison of theory with experiments. The second part will deal with questions of image formation and resolving power. The third part is devoted to discussion of technical details and the improvements of electron microscopes, whereas the fourth part will be devoted to the techniques used in electron microscopy and to different applications.

Due to necessary limitations of space, it might be worth while to point out forthwith a certain number of books and longer summaries which have appeared during the intervening period. The most remarkable is a book by v. Ardenne (1940 a) which, although restricted more or less to his own field of development, is clearly written and well documented. Another German publication, giving a very good cross-section of the work accomplished in a particular laboratory, is the *Jahrbuch der A.E.G. Forschungs-Institut* (1940) which has been republished with additions in 1942 by Ramsauer. Three more German books have appeared in the intervening period: Picht (1939), Ramsauer (1941), and Brüche and Recknagel (1941). These books, however, refer only occasionally to electron microscopy and, therefore, have less importance for the general development of the field. The English language publications during this period comprise one book by Burton and Kohl (1942) and one by Hawley (1945). Both books are addressed almost exclusively to the layman and do not contain any important technical information.

A certain number of summaries appeared in different publications, of which the following might be listed: v. Borries and Ruska, E. (1941), Anderson (1942), Marton (1943), Beischer (1943), Prebus (1944), Ruska, E. (1944), and a series of papers in a book, *Das Übermikroskop als Forschungsmittel*, which is at present unavailable in this country. In addition to these papers, bibliographies of electron microscopy have been compiled by C. Marton and S. Sass (1943–45).

§2. THEORY OF LENSES AND COMPARISON WITH EXPERIMENTAL DATA

Whereas the first years of electron optics saw the development of the approximate theory of relatively weak lenses, it has been recognized that the approximations far from satisfied the requirements of short-focus electron microscopy.

The next step, therefore, was to find analytical expressions for the lenses used in practice, to substitute them in the paraxial-ray equation and find the general solution of the resulting differential equation.

The most important papers about strong magnetic lenses were published by Glaser (1940a, 41) and Dosse (1941a–41c). Glaser first assumed that the measured field distribution $H(z)$ on the optical axis z can be approximated very closely by the expression

$$H(z) = \frac{H_0}{1+(z/a)^2} \qquad \ldots\ldots(1)$$

Here H_0 is the maximum field strength, at $z=0$, and a is that co-ordinate where $H(z)$ drops to $H_0/2$. $2a$ therefore may be called the half width of the field curve, as shown in figure 1.

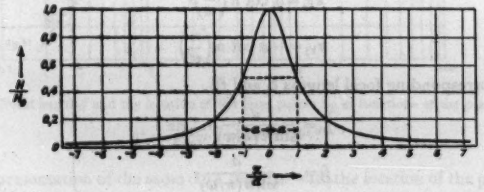


Figure 1. Field distribution along the optical axis of a magnetic lens of the field form $H(z) = \frac{H_0}{1+(z/a)^2}$. (Glaser, 1941.)

The paraxial-ray differential equation for a magnetic field of rotational symmetry is

$$\frac{d^2r(z)}{dz^2} + \frac{e}{8mV} H^2(z)r(z) = 0, \qquad \ldots\ldots(2)$$

where $r = r(z)$ is the equation for the paraxial-ray path of an electron in the magnetic field whose axial field-distribution is given by $H(z)$. e/m is the specific charge of the electron and V is the accelerating potential.

By substituting (1) in (2) we obtain

$$\frac{d^2r(z)}{dz^2} + \frac{eH_0^2}{8mV} \frac{1}{[1+(z/a)^2]^2} r(z) = 0. \qquad \ldots\ldots(3)$$

The general solution of this equation was given by Glaser. He obtained

$$r(z) = a\sqrt{1+(z/a)^2}\{C_1 \sin[\sqrt{1+k^2}\,\mathrm{arc\,cot}\,(z/a)] + C_2 \cos[\sqrt{1+k^2}\,\mathrm{arc\,cot}\,(z/a)]\}. \qquad \ldots\ldots(4)$$

where C_1 and C_2 are two arbitrary constants of integration and $k^2 = \frac{eH_0^2a^2}{8mV}$ is a parameter characterizing the lens strength. C_1 and C_2 can be determined so that the path satisfies any initial or boundary conditions.

This general solution of the paraxial-ray equation has remarkable properties. Although the familiar Listing construction of conjugated object and image elements has been used in electron optics in conjunction with weak lenses, it

was by no means sure that the same construction could be applied to strong lenses. In particular, considerable doubts were expressed about its validity in the case of image formation of objects which are no longer in a field-free region of the object space. It was thus very important when Glaser demonstrated that for this particular type of lens Newton's lens equation and magnification formula apply and that focal points and principal planes can be rigorously calculated, as in the light-optical case. For instance, the focal points can be determined in the following manner: C_1 and C_2 are chosen so that $r(z)$ is finite at $z=+\infty$ and $dr/dz=0$ for $z=+\infty$ (i.e., a parallel ray enters the lens, which begins at $z=+\infty$). The focal points are located at all values of z for which the solution has roots. The resulting expressions for the location of the focal points F_0 and F_1 are:

$$\left.\begin{aligned} z_{F_0} &= a \cot n\left(\frac{\pi}{w}\right), \\ z_{F_1} &= -a \cot n\left(\frac{\pi}{w}\right). \end{aligned}\right\} \quad \ldots\ldots(5)$$

and the corresponding focal lengths f_0 and f_1

$$\left.\begin{aligned} f_0 &= \frac{a}{\sin n\,(\pi/w)}, \\ f_1 &= -\frac{a}{\sin n\,(\pi/w)}, \end{aligned}\right\} \quad \ldots\ldots(6)$$

where

$$w=\sqrt{1+k^2}. \quad \ldots\ldots(7)$$

The positions of the principal points h_0, h_1 are determined by the relations

$$\left.\begin{aligned} z_{h_0} &= a \cot \frac{\pi}{2w}, \\ z_{h_1} &= -a \cot \frac{\pi}{2w}. \end{aligned}\right\} \quad \ldots\ldots(8)$$

Either w^2 or k^2 can be used to characterize the lens strength.

For this particular lens, both Newton's lens equation

$$(z_0-z_{F_0})(z_1-z_{F_1})=f_0 f_1, \quad \ldots\ldots(9)$$

and the magnification expression

$$M=(-1)^{n-1}\frac{f_0}{z_0-z_{F_0}}=(-1)^{n-1}\frac{z_1-z_{F_1}}{f_1} \quad \ldots\ldots(10)$$

are valid, although object and image are located within the lens.

If w has a value such that $N<w<N+1$, where $N, N+1$ are successive integers, the lens characterized by w will have N focal points. Consequently

$$0<\frac{eH_0^2a^2}{8mV}<3 \quad \text{or} \quad 1<w<2 \quad \ldots\ldots(11)$$

is the condition that single images exist. The shortest obtainable focal length is equal to a, the half width of the symmetrical field represented by (equation 1). Above the value $k^2=3$ multiple images occur with impaired image definition. For all practical purposes, electron microscopy is restricted to lenses contained within the range indicated by (11), although v. Ardenne (1941 c) proposes in one paper the use of the second-order focus for high-voltage microscopy.

The variation of the focal length and of the location of the focal points as functions of parameter k^2 is well illustrated by the diagram shown in figure 2.

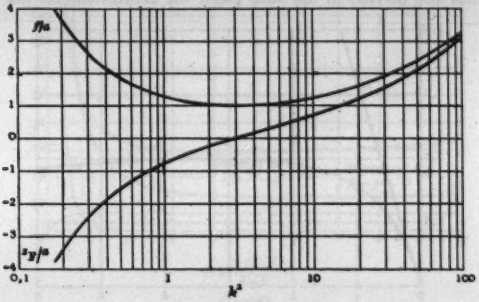


Figure 2. Focal length f and the location of the focal points z_F as functions of the parameter

$$k^2 = \frac{eH_0^2a^2}{8mV}.$$ (Glaser, 1941.)

Another representation of the same data, together with the location of the principal points, is given in figure 3.

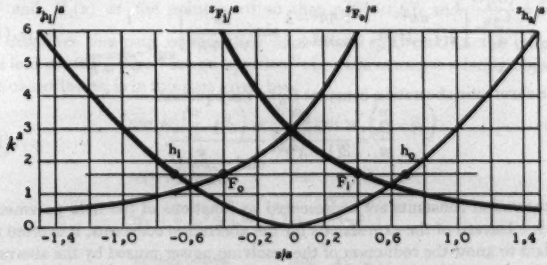


Figure 3. Location of the focal points and of the principal points as functions of the parameter k^2. (Glaser, 1941.)

It is interesting to compare the results thus obtained with the focal length calculated from the approximate expression derived by Busch (1927). This expression, commonly used earlier in the design of magnetic lenses,

$$\frac{1}{f} = \frac{e}{8mV}\int_{-\infty}^{+\infty} H^2\,dz, \qquad \ldots\ldots(12)$$

can be compared to the values obtained by Glaser by substituting for H^2 the expression (1). Figure 4 contains the calculated values for both equations, showing that the equation of Busch corresponds to an approximation for infinitely small values of the parameter k^2. Equation (12) is valid, therefore, only for very

weak lenses, and if applied to strong lenses, as used in electron microscopes ($1<k^2<2$), differences of 300 per cent and more can be obtained.

Glaser has also derived in the same paper the expressions for the spherical

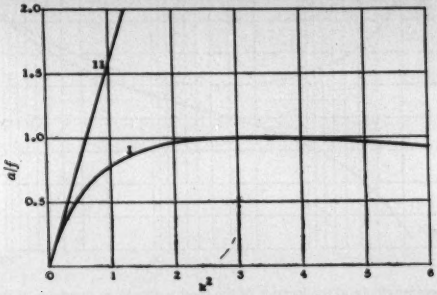


Figure 4. Refractive power a/f of a magnetic lens of the field form $H(z)=\frac{H_0}{1+(z/a)^2}$. Curve I is calculated from equation (6), curve II from equation (12).

and chromatic aberrations of the lens represented by equation (1). The spherical aberration constant C_s divided by the half width is given by

$$\frac{C_s}{a}=\left[\frac{\pi k^2}{4(1+k^2)^{3/2}}-\frac{1}{8}\frac{4k^2-3}{4k^2+3}\sin\frac{2\pi}{\sqrt{1+k^2}}\right]\frac{1}{\sin^4\frac{\pi}{\sqrt{1+k^2}}} \qquad \text{.....(13)}$$

and the chromatic aberration constant C_c by the equation

$$\frac{C_c}{a}=\frac{\pi k^2}{2(1+k^2)^{3/2}}\frac{1}{\sin^2\frac{\pi}{\sqrt{1+k^2}}}. \qquad \text{.....(14)}$$

Both aberration constants are represented as functions of the field parameter in figure 5. Instead of the expressions for the aberration constants, it is often more important to know the reduction of the resolving power caused by the aberration. The following equations give the relationship for both the spherical and the chromatic aberrations:

$$\delta_s=C_s\alpha^3, \qquad \text{.....(15)}$$

$$\delta_c=C_c\frac{\Delta V}{V}\alpha, \qquad \text{.....(16)}$$

where δ_s, δ_c are the radii of the circles of least confusion, produced by the aberrations, divided by the image magnification and α is the angular aperture of the lens.

A comparison of the theoretical results with experimental values was carried out by Dosse (1941 a–41 c). First Dosse showed that the fields most used in electron microscopy are asymmetrical. This is due to the fact that the pole-pieces of the magnetic lens are not identical in shape and the field distribution on the image side differs from that on the object side. A mathematical treatment

of such fields is possible by combining two bell-shaped field forms of different half widths a_1, a_2, and equal maximum field strength H_0.

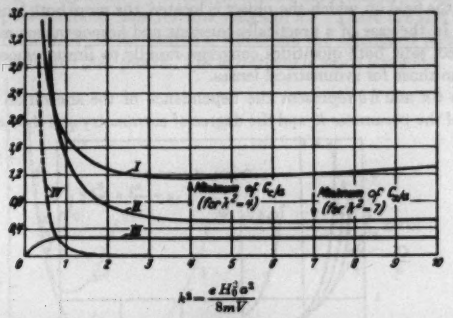


Figure 5.

I. Chromatic aberration constant C_c/a for high magnifications.
II. Spherical aberration constant C_s/a for high magnifications.
III. Spherical aberration constant C_s/a for optimum object position.
IV. Spherical aberration constant calculated in first approximation for a "weak lens". (Dosse, 1941 c.)

A path equation is then obtained by matching two solutions for the two fields $H_1(z)$ and $H_2(z)$ at the point $z=0$ so that $r_1(0)=r_2(0)$ and $\frac{dr_1(0)}{dz}=\frac{dr_2(0)}{dz}$. $r_1(z)$ describes the path on one side of $z=0$ and $r_2(z)$ that on the other side of the lens centre. The "image equation", i.e., the equation relating object and image co-ordinates, is in this case given by

$$\tan w_2\left(\frac{\pi}{2}-\phi_1\right)=\frac{a_1}{a_2}\frac{w_2}{w_1}\tan w_1\left(\frac{\pi}{2}-\phi_0\right), \qquad \text{.....(17)}$$

where

$$\left.\begin{aligned} z_0&=a_1\cot\phi_0, && \text{for } z\leqq 0;\\ w_1&=\sqrt{k_1^2+1}=\sqrt{1+\frac{eH_0^2a_1^2}{8mV}};\\ z_1&=a_2\cot\phi_1, && \text{for } z\geqq 0;\\ w_2&=\sqrt{k_2^2+1}=\sqrt{1+\frac{eH_0^2a_2^2}{8mV}}. \end{aligned}\right\} \qquad \text{.....(18)}$$

Equation (17) becomes identical to (9) if $w_1=w_2=w$. The magnification is given by

$$M=\frac{k_2}{k_1}\sqrt{\frac{k_1^2+1}{k_2^2+1}}\frac{\sin\phi_1}{\sin\phi_0}\frac{\sin[\sqrt{k_1^2+1}(\phi_1-\pi/2)]}{\sin[\sqrt{k_2^2+1}(\phi_0-\pi/2)]}. \qquad \text{.....(19)}$$

The focal points are given by

$$\left.\begin{aligned} \tan\left[\sqrt{k_2^2+1}\left(\frac{\pi}{2}-\phi_{F_1}\right)\right]&=-\frac{k_1}{k_2}\sqrt{\frac{k_2^2+1}{k_1^2+1}}\tan\sqrt{k_1^2+1}\frac{\pi}{2},\\ \tan\left[\sqrt{k_1^2+1}\left(\frac{\pi}{2}-\phi_{F_2}\right)\right]&=+\frac{k_2}{k_1}\sqrt{\frac{k_1^2+1}{k_2^2+1}}\tan\sqrt{k_2^2+1}\frac{\pi}{2}. \end{aligned}\right\} \qquad \text{.....(20)}$$

Glaser (1941) and Dosse (1941 a) investigated the influence of the asymmetry on the chromatic and spherical aberrations. This analysis shows that the steeper the side of the field on which the object is located, the more both lens aberrations increase. In the case of a practically constant and homogeneous magnetic field on the object side, both quantities converge rapidly to limits whose values are smaller than those for symmetrical lenses.

Figures 6*a* and 6*b* represent the dependence of the aberration constants as functions of the parameter k^2 and the degree of asymmetry q of the field.

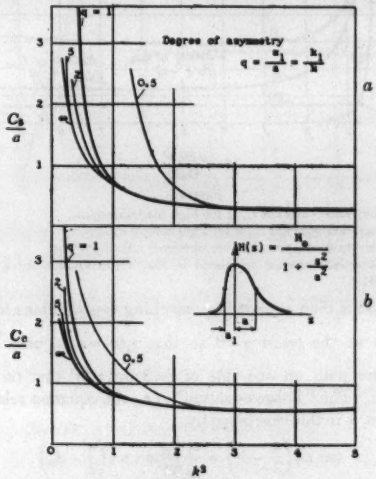


Figure 6.

(*a*) Spherical aberration constant as a function of k^2 for different degrees of field asymmetry.
(*b*) Chromatic aberration constant as a function of k^2 for different degrees of asymmetry. (Dosse, 1941 a.)

In a later paper Dosse (1941 c) compared in greater detail the results of theory with experiment. First he showed that, although the field distribution of saturated pole-pieces approximates relatively closely to the distribution given by equation (1), the field of unsaturated pole-pieces comes closer to the field represented by

$$H(z) = \frac{H_0}{\left(1 + \frac{z^2}{a^2}\right)^{\mu}}, \qquad \ldots\ldots(21)$$

where μ might be considerably different from 1. A complete theory of the magnetic lens, therefore, should take into account the variation of the exponent μ as a function of the magnetization, a task which might be very difficult to accomplish. For not too strong lenses, the discrepancy is small enough when comparing the measured and calculated focal lengths. Figure 7 represents such a comparison

of the measured focal length with theory for four different accelerating voltages. It shows a good agreement with the calculated focal-length for the case of unsaturated pole-pieces at least for low values of k^2. There are also a few points, indicated by crosses on the diagram, which are the calculated focal-lengths for the case of saturated pole-pieces. The agreement is much less good.

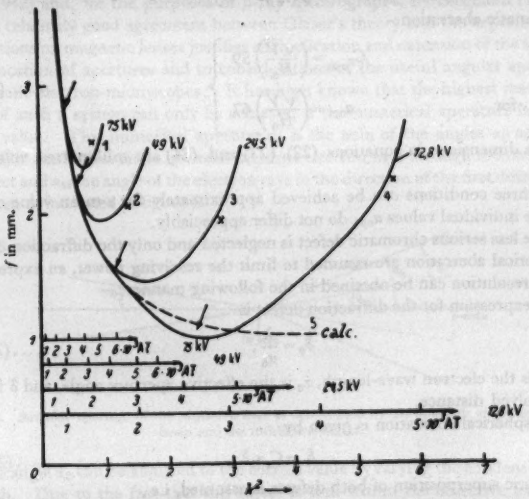


Figure 7. Experimental determination of the focal length f of a strong magnetic lens as a function of k^2. AT = ampere turns in the lens coil. The arrows indicate the beginning of saturation. (Dosse, 1941 c.)

When it comes to the investigation of the lens aberrations, experiments show that the observed values are relatively close to the calculated ones.

For practical design purposes Dosse (1941 c) derived some expressions which are much easier to use than the theoretical expressions. For high magnifications, the object is very near the focal point, and the aberration constants can be calculated for that point. The quantities C_s and C_c and the focal length f are functions of a^2 and k^2. If, therefore, H_0 and the accelerating potential V are kept constant, C_s, C_c and f become functions of a alone. It is then possible to determine the optimum value $a = a_{opt}$ for which each of these quantities reaches its minimum value. According to Dosse, the following expressions result:—

Focal length:

$$\left.\begin{aligned} f_{min} &= \left(\frac{\sqrt{V}}{H_0}\right) 85 \\ \text{for} \quad a_{opt} &= \left(\frac{\sqrt{V}}{H_0}\right) 60, \end{aligned}\right\} \qquad \ldots\ldots(22)$$

Spherical aberration:

$$\left.\begin{aligned} C_{s_{min}} &= \left(\frac{\sqrt{V}}{H_0}\right)35 \\ \text{for} \quad a_{opt} &= \left(\frac{\sqrt{V}}{H_0}\right)108. \end{aligned}\right\} \qquad \ldots\ldots(23)$$

Chromatic aberration:

$$\left.\begin{aligned} C_{c_{min}} &= \left(\frac{\sqrt{V}}{H_0}\right)59 \\ \text{for} \quad a_{opt} &= \left(\frac{\sqrt{V}}{H_0}\right)67. \end{aligned}\right\} \qquad \ldots\ldots(24)$$

(The dimensions in equations (22), (23) and (24) are millimetres, volts and gauss.)

All three conditions can be achieved approximately for a mean value of a_{opt} since the individual values a_{opt} do not differ appreciably.

If the less serious chromatic defect is neglected and only the diffraction defect and spherical aberration are assumed to limit the resolving power, an expression for best resolution can be obtained in the following manner:—

The expression for the diffraction defect is

$$\delta_0 = \frac{\lambda}{\alpha_0}, \qquad \ldots\ldots(25)$$

where λ is the electron wave-length, α_0 is the effective aperture angle, and δ is the least resolved distance.

The spherical aberration is given by

$$\delta_s = C_s \alpha_0^3. \qquad \ldots\ldots(26)$$

If quadratic superposition of both defects is assumed, i.e.,

$$\delta = \sqrt{\delta_0^2 + \delta_s^2} = \sqrt{\frac{\lambda^2}{\alpha_0^2} + \alpha_0^6 C_s^2}, \qquad \ldots\ldots(27)$$

then a minimum $\left(\frac{d\delta}{d\alpha_0} = 0\right)$ can be obtained for

$$\alpha_{0_{opt}} = 0{\cdot}872 \sqrt[4]{\frac{\lambda}{C}} \qquad \ldots\ldots(28)$$

δ_{min} is given by

$$\delta_{min} = 1{\cdot}32\, \lambda^{3/4} C_s^{1/4}. \qquad \ldots\ldots(29)$$

The wave-length of the electrons is related to their speed, as expressed in volts, by

$$\lambda = 1{\cdot}23 \times 10^{-6} V^{-1/2} \text{ [millimetre]}. \qquad \ldots\ldots(30)$$

Combining (28) and (29) with (30) and (23) gives

$$\delta_{min} = \frac{120}{\sqrt[4]{V H_0}} 10^{-6} \text{ [millimetre]}, \qquad \ldots\ldots(31)$$

for

$$\alpha_{0_{opt}} = 1{\cdot}2 \times 10^{-2} \sqrt[4]{\frac{H_0}{V}}, \qquad \ldots\ldots(32)$$

where V is in volts and H_0 is in gauss.

Brief reference will be made here to other papers dealing with the properties of magnetic lenses. Becker and Wallraff investigated experimentally in four papers (already quoted in Klemperer (1940)) the aberrations of relatively weak magnetic lenses at electron speeds corresponding to a range comprised between 4 and 20 ekv. Strong magnetic lenses have been treated theoretically by Svartholm (1942) and, for the purposes of β-ray spectrographs, by Siegbahn (1943).

The relatively good agreement between Glaser's theory and the experimental observations on magnetic lenses justifies an application and extension of the theory to the location of apertures and to considerations of the useful angular aperture in magnetic electron-microscopes. It has been known that the highest resolving power of such a system can only be achieved if the numerical aperature has the proper value. This numerical aperture α_0 is the sum of the angles α_C and α_D, where α_C is the largest angle between incident electron rays through the centre of the object and α_D the angle of the electron rays in the direction of the first diffraction maximum as indicated in figure 8.

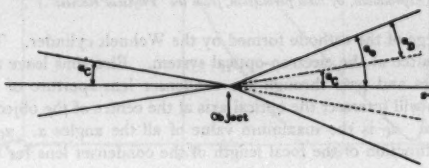


Figure 8. Angular aperture of the objective lens as determined by the aperture of the incoming beam and the diffraction defect.

The angle α_C can be adjusted to the correct value by varying the condenser lens strength. Due to the fact that the object is well within the objective lens, the value of α_C is not determined by the focusing action of the condenser lens alone. In microscopical work it has previously been assumed that the field of the objective lens, which extends beyond the specimen into the region between it and the condenser lens, is not strong enough to exert an appreciable converging effect on the illuminating beam of electrons (v. Borries and Ruska (1939c)). For high magnifications the specimen is very near the focal point, which for strong lenses is close to the lens centre ($k^2=3$, $z_{F_2}=z_{P_1}=0$). This means that the electrons coming from the condenser lens have to traverse part of the objective lens before reaching the object. The converging effect on the illuminating beam thus is far from negligible.

If it is assumed that the field forms of objective and condenser lenses are of the same type described by (1) and, furthermore, that a change in lens current does not alter the half width of the field curve but varies only the maximum field-strength H_0, α_C can be calculated in such a manner that the converging effect of the portion of the objective-lens between the specimen and the condenser is included, as demonstrated by Marton and Hutter (1944a).

If the distance between the condenser and the objective lens is large compared with the half widths of the field curves, an almost field-free plane can be assumed to exist midway between the two lenses. This assumption makes it possible to

compose the electron path through both lenses, using two separate solutions for each lens, and matching them in the field-free plane so that both their distance from the axis and their slopes are identical (see figure 9).

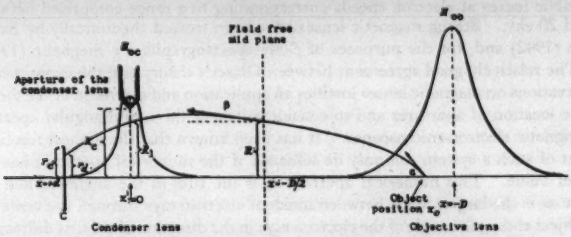


Figure 9. Path of an electron through the condenser and objective lens fields.
(*Reproduced, by kind permission, from the 'Physical Review'.*)

C is the image of the cathode formed by the Wehnelt cylinder. This image is the virtual source of the electron-optical system. Electrons leave this source at various angles and pass through the condenser lens aperture of radius R_c. These electrons will intersect the optical axis at the centre of the object, forming various angles α. α_C is the maximum value of all the angles α. α_C has been calculated as a function of the focal length of the condenser lens for the special conditions:

$$\left.\begin{array}{lll} z_0 = 0 \text{ (object at focus of the objective lens)}, & & \\ D = 420 \text{ mm.}, & d = 250 \text{ mm.}, & d_c = 3 \text{ mm.}, \\ d_0 = 1{\cdot}5 \text{ mm.}, & \gamma_c = 0{\cdot}05 \text{ mm.}, & R_c = 0{\cdot}4 \text{ mm.} \\ w_0 = 1{\cdot}5 \text{ mm.} & & \end{array}\right\} \quad \ldots\ldots(33)$$

α_C is plotted in figure 10 as a function of $1/f$ (curve I). For comparison the values of α_C calculated on the basis of formulae given previously by v. Borries and Ruska, E., are also plotted (curve II). It can be seen that it is not possible to decrease α_C below any prescribed limit by increasing the refractive power of the condenser lens. A definite minimum for α_C exists, and it is noteworthy that this is in accord with experimental results.

The physical aperture of the objective lens has to be of such a size as to allow only those rays to pass which leave the centre of the object forming an angle smaller or equal to α_0. The calculations of this aperture diameter were based on the assumptions that the electron paths are straight lines to the aperture edges, that the object is very near the focal point (high magnification), and that the aperture position is at the centre of the lens. Apertures determined by this method do not in fact have the intended ray-limiting effect. Marton and Hutter (1944) showed that optimum apertures can be obtained in the following manner. The electron path (equation (4)) reaches its maximum at some distance z_{opt} from the centre of the lens as indicated on figure 11. This distance is given by

$$\text{arc cot}\left(\frac{z_0}{a}\right) = \text{arc cot}\left(\frac{z_{opt}}{a}\right) + \frac{\pi}{w} - \left(\frac{1}{w}\right)\text{arc cot}\left(\frac{z_{opt}}{aw}\right) \quad \ldots\ldots(34)$$

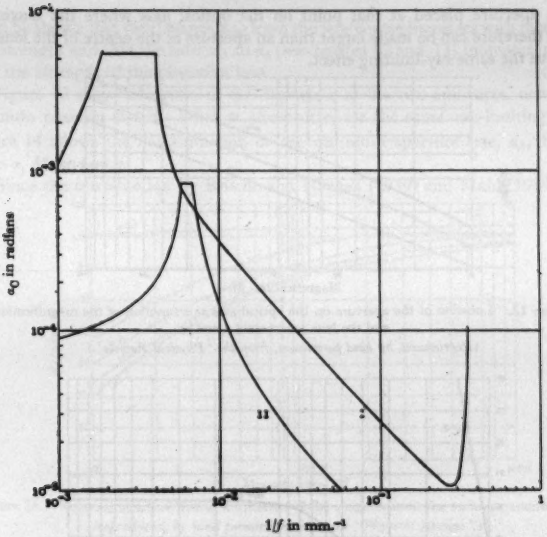


Figure 10. Angular aperture of the incoming beam as a function of the refractive power of the condenser lens. Curve I is as calculated by Marton and Hutter (1944 a), curve II as calculated on the basis of equations given by v. Borries and Ruska (1939 c).

as a function of z_{0}, the object position and the lens parameters a and w. Combining (34) with the formula for the magnification equation (10) (for $n=1$) it follows that

$$M=\frac{\sin\left[\text{arc cot}\,\frac{z_{opt}}{a}+\frac{\pi}{w}-\left(\frac{1}{w}\right)\text{arc cot}\left(\frac{z_{opt}}{aw}\right)\right]}{\sin\left[\left(\frac{1}{w}\right)\text{arc cot}\left(\frac{z_{opt}}{aw}\right)-\text{arc cot}\left(\frac{z_{opt}}{a}\right)\right]}. \quad \ldots\ldots(35)$$

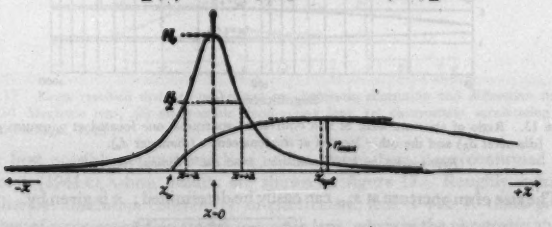


Figure 11. Optimum conditions for the location of a ray-limiting aperture.

Equation (35) is plotted in figure 12. z_{opt} can be determined from this plot for various M, a, and w.

An aperture placed at that point on the optical axis where the maximum occurs therefore can be made larger than an aperture at the centre of the lens and still have the same ray-limiting effect.

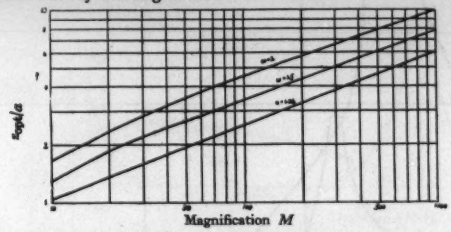


Figure 12. Location of the aperture on the optical axis as a function of the magnification and the lens parameters a and w.
(*Reproduced, by kind permission, from the 'Physical Review'.*)

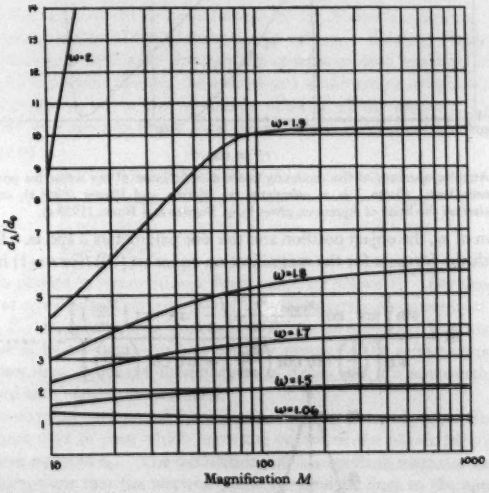


Figure 13. Ratio of aperture sizes of two equivalent apertures, one located at optimum position (diameter d_A) and the other located at the lens centre (diameter d_0).
(*Reproduced, by kind permission, from the 'Physical Review'.*)

The size of an aperture at z_{opt} can easily be determined; it is given by

$$\frac{r_A}{a\tan\alpha}=a\sqrt{\frac{\left(1+\frac{z_{opt}}{a}\right)^2\left(1+\frac{z_0}{a}\right)^2}{w^2+\left(\frac{z_{opt}}{a}\right)^2}}. \qquad \ldots\ldots(36)$$

It can be shown that the distance z_{opt} from the lens centre increases with increasing lens strength and that the ratio d_A to d_0 (see figures 12 and 13) increases rapidly with the strength of the objective lens.

Figure 13 shows the ratio of the diameters of the two apertures, one at the optimum position and the other at the centre, for the same ray-limiting effect. Figure 14 allows the determination of the optimum aperture size, d_A, for any given α, M, w, and a.

Since the introduction by Brüche and Häagen (1939) and Mahl (1939, 40 a)

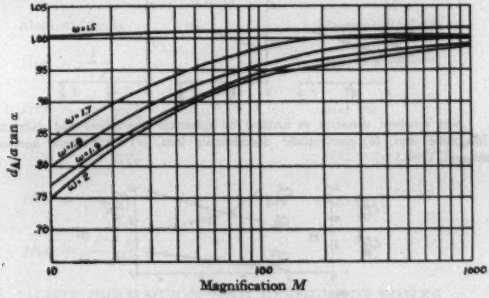


Figure 14. Reduced aperture size as a function of the magnification for various parameters w. (*Reproduced, by kind permission, from the 'Physical Review'.*)

of the electrostatic electron-microscope, some attention has been devoted to the properties of strong electrostatic lenses.

The first comparative calculations of lens aberrations were carried out by Recknagel (1940). Without going into detail, Recknagel came to the conclusions represented in figures 15 and 16, showing that the best magnetic lens is superior

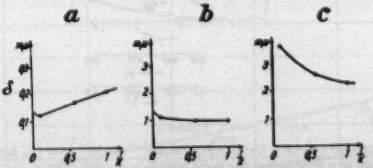


Figure 15. Least resolved distance as limited by chromatic aberration and diffraction defect: (*a*) Magnetic lens, (*b*) electrostatic decelerating lens, (*c*) electrostatic accelerating lens. (Recknagel, 1940.)

to the best possible electrostatic lens. The results have been confirmed later by Dosse (1941 c), whose results are shown on figure 17. Roughly speaking, the spherical aberration constant of an electrostatic lens is 4 to 10 times greater than that of a corresponding strong magnetic lens, whereas the chromatic aberration is 4 to 6 times greater than that of the magnetic lens.

In spite of that limitation, for practical purposes the electrostatic lens can often be used, as demonstrated by Mahl (1940 c), v. Ardenne (1940 a), and by Bachman and Ramo (1942).

Complete and rigorous calculation of an electrostatic lens, as Glaser and Dosse did for the magnetic lenses, exists at present only for the potential function $\phi(z) = \phi_0 e^{\pm k \arctan z}$ (Hutter 1945a). As this lens does not belong in the class used in electron microscopy, its discussion is omitted here.

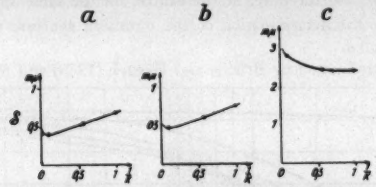


Figure 16. Least resolved distance as limited by spherical aberration and diffraction defect: (*a*) Magnetic lens, (*b*) electrostatic decelerating lens, (*c*) electrostatic accelerating lens. (Recknagel, 1940.)

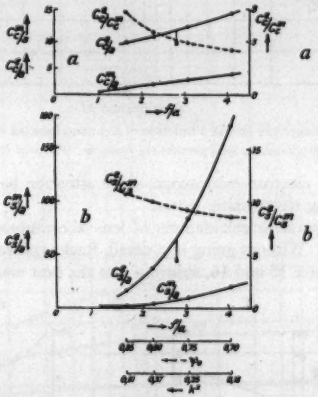


Figure 17.

(*a*) Chromatic aberration constant for an electrostatic "single" lens (*e*), for a magnetic lens (*m*), and their ratio as a function of the focal length.

(*b*) Spherical aberration constants for the same lenses. (Dosse, 1941 c.)

The discussion about relative merits of electrostatic and magnetic lenses has been particularly sharp in Germany. For further reference on this discussion we refer here to the papers and books published by v. Borries and Ruska, E. (1940), Ramsauer (1940), Brüche (1941), and Ruska, E. (1942). This last paper contains a rather interesting diagrammatic representation of the minimum focal lengths of electron lenses at various electron-beam speeds, which we reproduce in figure 18. It also contains some considerations of an economic nature relating to the construction of magnetic electron microscopes.

A further interesting point is to investigate which are the field distributions for which Newton's image equations apply, and the traditional Listings construction can be used. Two papers have been devoted to the subject, one by Glaser and Lammel (1941)and more recently one by Hutter (1945 b). It can be shown that there exist a few fields which, in spite of producing "strong" lenses, fulfil the above requirements. Some of the physically realizable field functions are listed by Hutter. Table 1 reproduces his results:—

Table 1

Magnetic fields	Electrostatic fields
$H(z)=\frac{1}{1+z^2}$	$\phi(z)=e^{c\arctan z}$
$H(z)=\frac{1}{(1+z^2)^{3/2}}$	$\phi(z)=e^{c\frac{1}{\sqrt{1/z^2+1}}}$
$H(z)=\frac{z}{(1+z^2)^{3/2}}$	$\phi(z)=c_1 e^{c_2\frac{1}{\sqrt{z^2+1}}}$
$H(z)=\frac{z}{(1+z^2)^2}$	$\phi(z)=c_1 e^{c_2\frac{1}{1+z^2}}$
$H(z)=\frac{\sqrt{z^2+1}}{(1+z^2)^2}$	

§3. IMAGE FORMATION AND RESOLVING POWER

The primary phenomenon in image formation of a transmission type of electron microscope is diffraction, as in the case of a light instrument. The diffraction phenomena, however, are complicated by scattering of electrons in the substance of the specimen and by energy losses. The resulting effects cannot be represented in a strictly mathematical form on account of the superposition of the different effects.

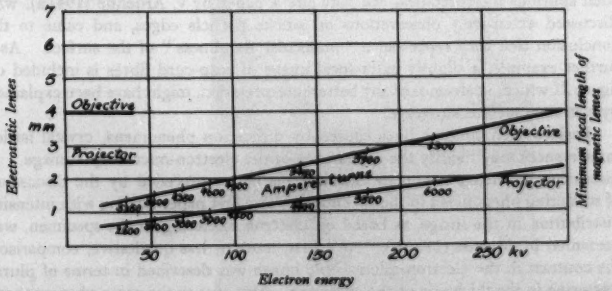


Figure 18. Focal lengths of electron lenses as functions of the electron energy. (Ruska, 1942.)

It was recognized early in the history of electron microscopy that diffraction plays a very important rôle. To some extent its rôle should have been discussed in the preceding chapter, under the heading of lens properties, because the diffraction defect influences the performance of the lens. It was briefly mentioned

that, on account of diffraction, the smallest separation which can be resolved is inversely proportional to the diameter of the physical aperture, in close similarity to light optics. It had to be demonstrated experimentally, however, that the observed limitation is truly due to diffraction. The best demonstration of diffraction phenomena in electron microscopes is due to Boersch (1945 b).

Since the beginning of high-power electron microscopy, wide contour lines following the outline of particles have often been observed. In 1939 two interpretations of this phenomenon were given, one by v. Borries and Ruska, E. (1939 a) and another by Hillier (1939). Both interpretations were highly qualitative. Whereas v. Borries and Ruska attributed these contour lines to variations in the spherical aberration of the objective for different regions of the specimen, Hillier explained such contour lines at the line of separation of two areas of the specimen, having widely different atomic mass thicknesses, as due to chromatic aberration. A year later Boersch (1940) demonstrated with an electron-shadow microscope (Boersch, 1939) the presence of Fresnel diffraction fringes, while at about the same time Hillier (1940) interpreted a series of fringes observed on supposedly sharply focused images as a Fresnel diffraction phenomenon. The same phenomenon was then investigated much more in detail by Boersch (1943 a, 44) and by Ruska, E. (1943). The most important results of these investigations are that, if the numerical aperture of the optical system be sufficiently reduced, Fresnel diffraction fringes can be observed in out-of-focus images. Some of the most beautiful observations of Boersch are reproduced on figures 19 and 20. They can be relatively easily distinguished from crystal lattice interferences by their position and by their dependence on the focusing of the image. They are due to diffraction on the edges of particles which show considerable contrast, as shown by a demonstration through dark-field illumination (Boersch, 1943 a). The demonstration of such effects became very important because it helped to eliminate different spurious interpretations of observed phenomena. As an example of such spurious interpretation, we may cite a paper by v. Ardenne (1941 a), who discussed extensively observations on smoke particle edges, and came to the conclusion that they represent a " molecular roughness " of the surface. As a further example, a slightly extra-focal image of soap-curd fibrès is included on figure 21 which, in absence of any better interpretation, might have been explained by a strange surface structure.

Besides such contour lines caused by diffraction phenomena, crystal lattice interferences may modify the appearance of the electron-microscopic image. A discussion of such phenomena, however, is better preceded by the discussion of scattering phenomena in the specimen. The first paper dealing with intensity distribution in the image, as based on electron scattering in the specimen, was presented by Marton (1936 b). In a first, more or less qualitative, comparison the contrast in the electron-microscopic image was described in terms of plural scattering in the thickness of the specimen. Also, due to the same phenomenon, a distinction was made between resolving power as due to diffraction phenomena and " depth resolving power " as due to limitations of contrast. A natural consequence of these effects is that a specimen, placed on a supporting layer, has to face the objective lens, and not be viewed through the support, for a full use of the resolving power of the instrument. Two years later v. Ardenne (1938 c)

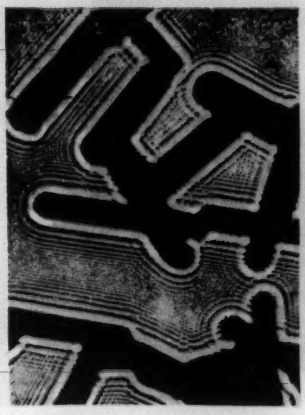


Figure 19. Fresnel diffraction on zinc oxide crystals. (Boersch, 1943 d.)

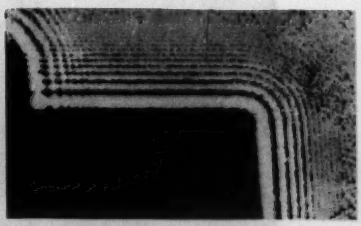


Figure 20. Fresnel diffraction in extra focal observation. (Boersch, 1943 d.)

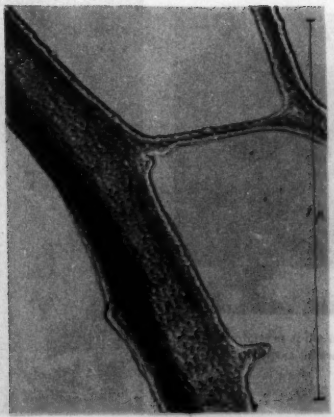


Figure 21. Spurious " surface structure " of soap-curd fibres in extra-focal image.

To face page 220

Figure 22. Crystal interference fringes on molybdenum oxide.

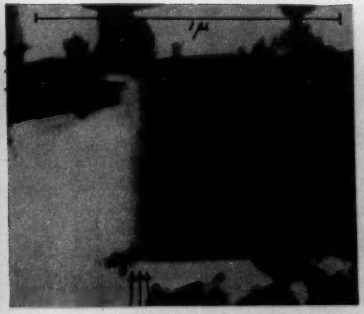


Figure 23. Magnesium oxide crystal with interference fringes parallel to one edge.

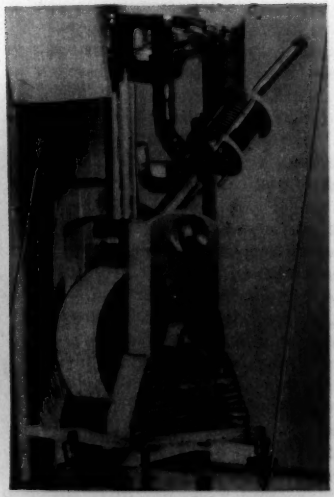


Figure 37. Electron microspectroscope.

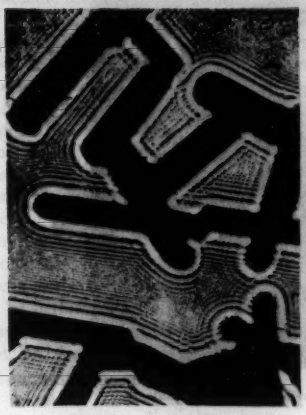


Figure 19. Fresnel diffraction on zinc oxide crystals. (Boersch, 1943 d.)

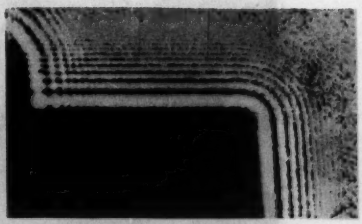


Figure 20. Fresnel diffraction in extra focal observation. (Boersch, 1943 d.)

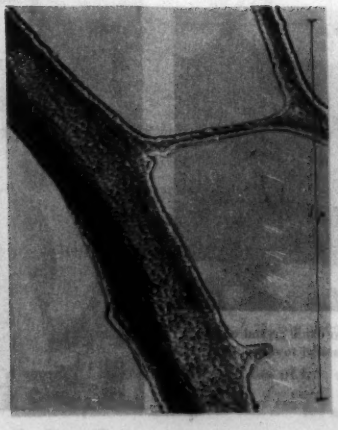


Figure 21. Spurious "surface structure" of soap-curd fibres in extra-focal image.

To face page 220

Figure 22. Crystal interference fringes on molybdenum oxide.

Figure 23. Magnesium oxide crystal with interference fringes parallel to one edge.

Figure 37. Electron microspectroscope.

introduced an expression for the limiting of resolving power due to scattering. Another paper (v. Ardenne, 1938 a) contained a computation of the limitation of resolving power due to all aberrations, and an expression was given for the best attainable resolving power by quadratic summation of all the factors.

The same type of calculation was the object of an investigation by Marton and Schiff (1941). As most specimens of electron microscopy are relatively thin, they give rise mostly only to single scattering. The theory of single scattering for relatively low electron-energies (comprised between 20 and 100 ekv.) and for low angles of scattering is not very extensively developed. Three types of scattering had to be considered: elastic and inelastic scattering by the atom or ion, and scattering by the conduction (free) electrons in the case of a metal. Since the elastic scattering may be regarded as a nuclear or Rutherford scattering, made finite at small angles by the screening effect of the atomic electrons, it could be assumed that results based on isolated neutral atoms can be used, and that possible ionization in the solid can be disregarded. Such an approximation could not be used in the case of inelastic scattering because this latter is accompanied by excitation of the atom. The calculations for this last case had to be performed for free ions where metals or salts of the alkalis or halogens were involved. Without going into detail, the following expressions were obtained for the different scattering cross-sections:—

Elastic scattering cross-section:

$$\left.\begin{aligned} \sigma_e &= A\left(\frac{\hbar}{mv}\right)^2 && \text{for } \alpha<\theta_1, \\ \sigma_e &= 4\pi\left(\frac{Ze^2}{mv^2\gamma\alpha}\right)^2 && \text{for } \alpha>\theta_1. \end{aligned}\right\} \quad \ldots\ldots(37)$$

In elastic scattering cross-section:

$$\left.\begin{aligned} \sigma_i &= B\left(\frac{\hbar}{mv}\right)^2 \ln\frac{2}{\theta_2} && \text{for } \alpha<\theta_2, \\ \sigma_i &= B\left(\frac{\hbar}{mv}\right)^2 \ln\frac{2\theta_2}{\alpha^2} && \text{for } \theta_2<\alpha<\theta_3. \end{aligned}\right\} \quad \ldots\ldots(38)$$

Free-electron scattering cross-section:

$$\left.\begin{aligned} \sigma_f &= 8\pi\nu\left(\frac{e^2}{mv^2\gamma}\right)^2\left(\frac{3}{4\theta_4\alpha}-\frac{3}{8\theta_4^2}+\frac{\alpha}{16\theta_4^3}\right) && \text{for } \alpha<2\theta_4, \\ \sigma_f &= 4\pi\nu\left(\frac{e}{mv^2\gamma\alpha}\right)^2 && \text{for } \alpha>2\theta_4. \end{aligned}\right\} \quad \ldots\ldots(39)$$

In the equations (37), (38), (39) $\hbar$ is Planck's constant divided by 2π, α is the half-angle of aperture of the objective lens of the microscope, $\gamma=\frac{1}{\sqrt{1-v^2/c^2}}$, θ_1 is the "cut-off" angle at which screening of the nucleus by the inner electrons plays an important rôle, $\theta_2=I/mv^2\gamma^2$, $\theta_3=\left(\frac{2I}{mv^2\gamma^2}\right)^{1/2}$, where I is the mean excitation energy of the atom, $\theta_4=\frac{2\pi\hbar}{mv\gamma}\sqrt[3]{\frac{3\nu N}{8\pi}}$ where, ν is the number of free electrons per atom and N is the number of atoms per unit volume, $A=22\,Z^{4/3}$ for all except the

lightest atoms and B is a dimensionless number which has been calculated using Slater wave functions. A, B and I are tabulated below for several cases of practical interest.

Table II. Parameters appearing in equations (37), (38), (39)
A and B are dimensionless numbers and I is given in ev.

	A	B	I		A	B	I
1 H	7·3'	13·3	13·5	17 Cl^-	930	76	77
2 He	10·6	8·8	38	18 A	875	55	110
3 Li^+	60	3·4	98	19 K^+	1110	43	140
6 C	246	35	44	20 Ca^{++}	1200	33	180
7 N	229	28	63	29 Cu^+	1950	91	130
8 O	209	23	83	35 Br^-	2520	120	130
9 F^-	274	26	91	37 Rb^+	2700	88	170
10 Ne	215	18	130	47 Ag^+	3720	200	170
11 Na^+	316	13	180	51 Sb	4150	220	180
13 Al^{+++}	547	7·5	300	53 I^-	4370	180	150
15 P	840	80	63	55 Cs^+	4600	120	230
16 S	846	70	74				

The total scattering cross-section S for scattering outside the angle limited by the objective aperture is given by

$$S=\sigma_e+\sigma_i+\sigma_f. \qquad \ldots\ldots(40)$$

The thickness x of an object, producing only single scattering, is given, therefore, by

$$x=\frac{1}{SN}\ln\frac{t_0}{t_1}, \qquad \ldots\ldots(41)$$

if proportionality is assumed between the transmissions t_1, t_0 of the photographic negative and electron-beam intensities reaching the emulsion. The index 1 refers to the beam passing through the specimen (+ support), the index 0 to the unscattered beam or scattered by the support alone. If the specimen is composed of different chemical elements, additivity of the atomic (or ionic) cross-sections is assumed and the thickness becomes

$$x=\frac{1}{\Sigma_i S_i N_i}\ln\frac{t_0}{t_1}, \qquad \ldots\ldots(42)$$

where $N_i=\dfrac{N_0 p_i n \rho}{M}$ with N_0 = Avogadro's number, p_i the fraction of atoms of type i in the molecule, n the total number of atoms in one molecule ($p_i n$ is therefore the number of type i atoms in one molecule), ρ the density and M the molecular weight. Thus for the calculation of thickness a knowledge of the chemical composition of the specimen as well as the density and of the molecular weight is necessary.

Such a calculation can be used for an estimation of the thickness of many specimens. It does not take into account, however, selective reflections produced on crystal-lattice planes. That such reflections can alter completely the intensity distribution was first demonstrated by v. Borries and Ruska (1940). They

observed that the appearance of chromic-oxide particles changed very considerably with slight changes of the angle of the incident beam. The same type of observation was repeated later by different observers. Another interesting group of observations belonging to the same class showed dark and bright fringes across the images of crystalline particles, changing position with changing angle of incidence. Figure 22 shows, for instance, such " interference fringes " on molybdenum oxide crystals, and figure 23 on cubic crystals of magnesium oxide. That these observations are all due to crystal-lattice interferences has been demonstrated by Boersch (1942 c), Hillier and Baker (1943), Kossel (1943), and Heidenreich and Sturkey (1945).

A qualitative explanation of these crystal-lattice interferences is illustrated on figure 24. If an incident electron beam I_0, which is supposed to be a parallel

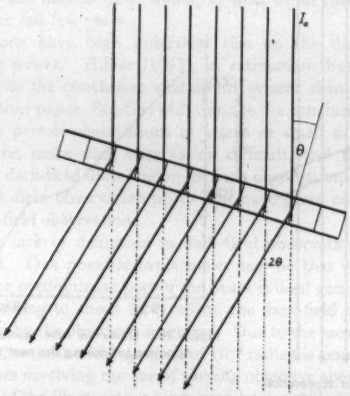


Figure 24. Schematic representation of Bragg reflection on crystal lamellae.

beam, falls upon a thin crystal, oriented at the Bragg angle θ, then the reflected beam on the crystal-lattice planes leaves the crystal at an angle 2θ. In the case of the most common type of electron-microscope observation, the objective aperture intercepts all rays leaving the specimen at an angle greater than $\alpha_0 \simeq 10^{-3}$ radian. A crystal observed under those conditions will, therefore, appear dark. By removing the objective aperture from an electron microscope and by slightly defocusing the instrument, two images can be observed. One, a dark image, corresponds to the first kind of observation with a small aperture, and the second, an offset, bright image, contains the reflected beam.

In the case of observation of fringes across thin crystal lamellae, it has to be assumed that the crystal is slightly warped or bent, as illustrated on figure 25. A small region of the crystal can present itself again at the proper orientation and produce a reflection of the beam beyond the solid angle produced by the objective aperture. Again, by removing the objective aperture, bright fringes appear in the image. The position of these bright fringes moves across the image with

varying focus with respect to the dark fringes. Wandering of the fringes across the lamellae can be observed directly on the fluorescent screen of the electron microscope with very intense beams. Such intense electron beams produce visible further warping of the crystals by thermal effects, or by charging or by both.

In the case of the fringes which are parallel to the edge of a cubic crystal, like the one shown on figure 23, Kossel (1943) attempted first to apply the dynamical theory of diffraction. By assuming that the reflections originated on a (220) plane, he calculated values of the "characteristic" crystal thickness D_0. His calculations give a rather considerable deviation from the Bragg angle. Later, Heidenreich and Sturkey (1945) recalculated the spacings and found

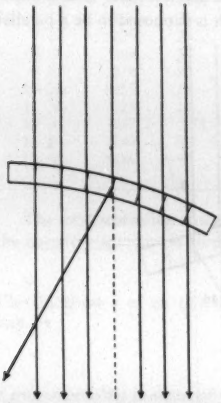


Figure 25. Schematic representation of the production of interference fringes on bent crystals.

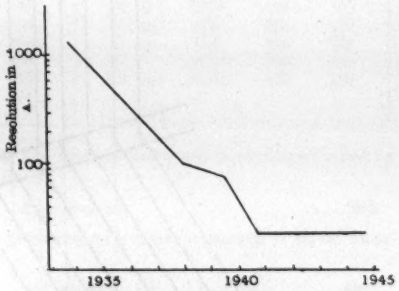


Figure 26. Observed best resolution of electron microscopes during the past twelve years.

considerably better agreement by assuming that the reflections originated on a (200) plane.

All of these observations and calculations show that the electron microscope can give a considerable amount of information on the properties of crystalline materials. The observations are still rather scarce, and much more material is needed. Heidenreich and Sturkey (1945) in particular propose a series of observations with varying illuminating apertures, varying objective apertures and different orientations of the crystals. That such an extended programme of investigation of crystalline materials is highly needed can be further illustrated by pointing to some obvious misinterpretations in earlier works, from which, in particular, another paper by v. Ardenne may be quoted (v. Ardenne and Beischer, 1940; reproduced also on pp. 302–303 of his book *Electronen-Übermikroskopie*). The "molecular chains" of about 30 A. thickness indicated on figure 285 of the book, are probably to be reinterpreted in view of our present knowledge, as an "anomalous" intensity-distribution on a small crystal.

A considerable amount of discussion was devoted in meetings of the Electron Microscope Society of America to proposed criteria and definition of the resolving power in electron microscopy. At the date of writing this *Report*, no agreement has been reached, and it is often found that different papers use different criteria for judging the quality of the image. These discrepancies are not so suprising if we consider the lack of agreement sometimes found in defining what constitutes the limit of resolution in light-microscope images. By sampling, however, all the papers published on electron microscopy, one comes to the conclusion that the best results achieved do not go beyond 20 A. In this respect it is interesting to compare the achievements of electron microscopy. Figure 26 represents the best resolving power as a function of time. It shows that apparently we have come to a more or less natural limit which, in spite of all efforts, has not been improved during the last few years.

Some calculations have been published also on the theoretical limit of attainable resolving-power. Hillier (1941), in estimations based on diffraction defect alone, came to the conclusion that atoms greater than $Z=25$ should be discernible. In a later paper, Schiff (1942) came to the conclusion that scattering phenomena should permit observations of atoms as small as $Z=7$. Contrast limitations, however; make such observation difficult, and that is why Schiff proposes the use of dark-field illumination for such observations. Unfortunately, all attempts at dark-field observation up to now have given considerably poorer results than bright-field observation.

The reason for lack of definition in dark-field observation has never been properly explained. One possible explanation may be that whereas in bright-field observation the predominant part of the beam is used generally at numerical apertures corresponding to about 10^{-5} radian, the dark-field image fills out the full opening offered by the objective aperture. Due to the location of such apertures, the assumed low numerical aperture of 10^{-3} radian is generally not fulfilled. Further experiments involving the use of varying objective apertures and varying angle of incidence of the illuminating beam may be desirable.

A very interesting contribution to the problem of resolving power was given by Glaser (1943) in a paper entitled *Image Formation and Resolving Power of the Electron Microscope from the Point of View of Wave Mechanics*. Starting with a criticism of earlier attempts for the determination of the resolving power by quadratic summation (as shown, for instance, in equation (27) or by other similar methods), he demonstrates that for the apertures most useful in practice it is senseless to speak of spherical aberration and of diffraction defect as separate phenomena. In the presence of spherical aberration (both for light and for electron optics) the image is no more a product of interference of spherical diffraction waves, but of aspherical ones. It thus becomes necessary to calculate wave-mechanically the electron intensity in the image plane as a result of such aspherical waves. From the different assumptions needed for such a calculation, there is one which seems to this reviewer a little doubtful: for the intensity-distribution function in the object plane Glaser assumes the classical Rutherford equation for multiple scattering, although he recognizes that, for most specimens, it would not apply. The suprising result of the whole calculation, which we have

to omit here, due to lack of space, is a justification, within a slight numerical factor, of equation (29). Glaser obtains for the resolving power

$$\delta_{\min} = 0{\cdot}56\,\lambda \sqrt[4]{\frac{C_s}{\lambda}} \qquad \ldots\ldots(43)$$

and for the aperture

$$\alpha_{\text{opt}} = 1{\cdot}13 \sqrt[4]{\frac{C_s}{\lambda}}. \qquad \ldots\ldots(44)$$

An interesting aspect of the image formation in electron microscopy is the correlation between the composition of the specimen and the observed image. We have seen earlier in this section that the scattering calculations alone do not answer this question completely, particularly in the case of specimens where the chemical composition is unknown. Also, such considerations do not apply in the case of crystalline specimens. Electron-diffraction experiments cannot solve the problem completely either, since they do not give generally a pattern which can be easily interpreted. A way out of these difficulties was opened up when Ruthemann (1941, 42) observed that characteristic energy losses at x-ray levels can be observed when a beam is passing through thin layers of matter. Since the discovery of Ruthemann that such observations are possible, they have been extended to the observation of very small specimens by Hillier (1943), by Hillier and Baker (1944), and, by adapting the method to electron microscopes, by Marton (1944 b). The instruments used for observation of such velocity losses will be described in the following section. It should now be pointed out, however, that besides rendering possible the determination of the chemical composition of specimens, the combination of a velocity analyser with an electron microscope, as used by Marton, may open the way to a quantitative investigation of the chromatic aberration of electron lenses, which up to now has been rather inaccessible to direct measurement.

§ 4. ELECTRON-MICROSCOPE DESIGN

The last *Report* in this series (Klemperer, 1940) described the earliest commercial instrument developed at the Siemens Works (v. Borries and Ruska, E., 1939 b). Also a brief description of an electrostatic instrument constructed at the A.E.G. Company (Brüche and Haagen, 1939; Mahl, 1939) was given. A number of other constructions have marked the intervening time, of which several will be described in the chronological order of their publication.

The instrument described by v. Ardenne (1940 a–40 c) was called by its designer a " universal electron microscope for bright field, dark field and stereo operation ". A cross-section drawing of this instrument is given in figure 27. Built as a very versatile instrument, its lens elements could be easily exchanged for other types of lenses. For instance, the magnetic lenses could be exchanged for electrostatic ones without any major changes. The magnetic lenses followed more or less the general construction principles of earlier designs, with particular attention, however, to the location of the apertures. Coils were provided with water-cooling. A considerable amount of detail is given about such constructional features in v. Ardenne's book (1940 a). One of the most interesting characteristics of this instrument was the introduction of an object-tilting device for stereoscopic electron-micrography. Following a suggestion made earlier by Ruska, a device,

shown in figure 28, was provided for taking the image of an object once at a certain angle of tilt, and the second image by tilting the object in the opposite direction.

The power supplies of this instrument follow conventional lines, with a

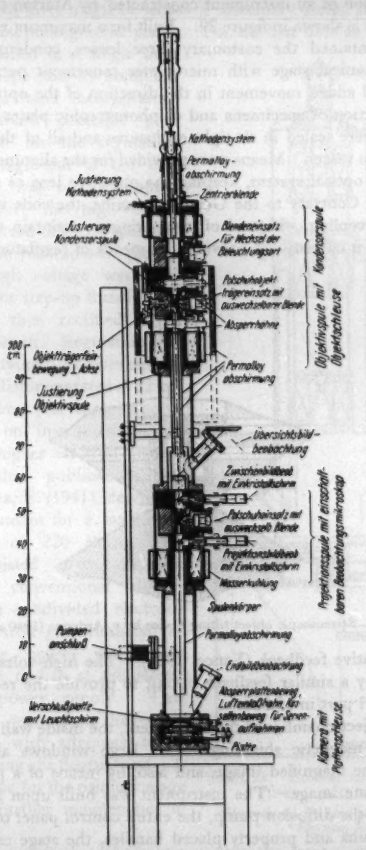


Figure 27. "Universal" electron microscope of v. Ardenne (1940 a, 40 b).

device for controlling manually the constancy of the high voltage. This control consists of a cathode-ray tube operated from the same accelerating voltage as the electron-microscope beam. The beam of this cathode-ray tube was deflected

through 90° by a constant magnetic field and the variations of deflection observed on a fluorescent screen. The high voltage was adjusted manually to give constant deflection of the control beam.

A cross-section of an instrument constructed by Marton (1940) (also called R.C.A. Type A) is shown in figure 29. Built for a maximum electron-velocity of 100 ekv., it contained the customary three lenses, condenser, objective and projector, mechanical stage with micrometer movement perpendicular to the optical axis, and added movement in the direction of the optical axis, air locks for the introduction of specimens and of photographic plates into the vacuum. The lens coils were sealed in air-tight enclosures and all of them housed inside the main vacuum vessel. Means were provided for the alignment of the different elements of the optical system, keeping the objective lens as the fixed point of the alignment. Contrary to the German practice, the coils were not provided with any water-cooling. Instead of attempting to maintain the coil resistance constant, constant current was provided by means of regulators operated on the

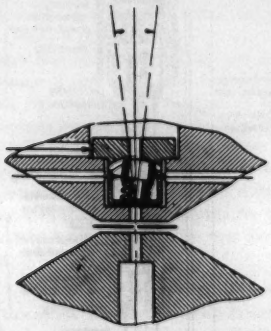


Figure 28. Stereoscopic object tilting device by v. Ardenne (1940 a).

principle of negative feedback (Vance, 1941). The high-voltage power supply was controlled by a similar feedback circuit to provide the required constancy of the voltage of 1 part in 15,000.

Below the object-chamber of the instrument, the inside wall of the main tube was lined with magnetic shielding. Two large windows allowed binocular observation of the magnified image, and also, by means of a periscope system, of the intermediate image. The instrument was built upon a metal pedestal, housing, besides the diffusion pump, the entire control panel of the instrument. By means of chains and properly placed handles, the stage could be operated from the same position of the observer, from where the electrical controls were operated and both the highly magnified and the intermediate images could be observed. Experiments carried out with this instrument showed that proper magnetic deflection of the incident (condenser) beam can be successfully used for the production of stereoscopic images (Marton and Schiff, 1941).

In 1941 the R.C.A. commercial electron microscope (called Type B) was announced (Hillier and Vance, 1941; Zworykin, Hillier and Vance, 1941a) (figure 30). The instrument was constructed for a maximum velocity of 60 ekv. It was provided with three magnetic lenses of customary design and with air locks for the specimen and photographic plates. For the first time an instrument was built which combined in a single unit both the microscope and its complete power supply, this latter incorporating the negative-feedback principle previously described for the regulation both of high voltage and of the coil currents. There was one significant difference however, the high-voltage power supply, instead of being operated on 60 cycles, employed radio-frequency power. The high voltage was produced in a resonant step-up transformer of high Q and then rectified in a voltage-doubler circuit. Regulation was achieved by modifying the power output of the master oscillator operated at 32 kc.

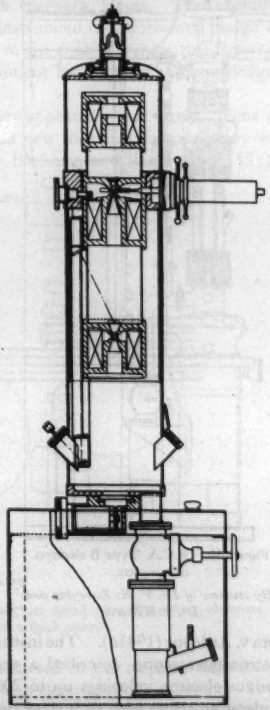


Figure 29. Electron microscope by Marton (1940). (R.C.A. Type A.)
(*Reproduced, by kind permission, from the 'Physical Review'.*)

During the same year three reports were published on instruments built for considerably higher electron-speeds. The first of thse publications, by Müller and Ruska, E. (1941), described a magnetic instrument for a maximum electron-velocity of 220 ekv. This instrument consisted of a Siemens microscope of conventional design provided with a subdivided electron-gun. The electron gun consisted of two accelerating stages with each accelerating stage tied in at the proper place on the power supply. This power supply was of a half-wave rectifying type, with adequate filter-circuit for reducing the fluctuations. A magnetic regulator for the primary stage was added for further maintenance of the constancy.

This first demonstration of electron-microscope images at higher velocities proved the value of such observations. Even in the first suggestion of higher-voltage electron-microscopy, Marton (1939 a) pointed out that such instruments will open up the way to the observation of much thicker specimens than had previously been possible with the lower voltage instruments. The main accent in considering higher velocities lies always in the greater penetration and not in

increased resolving power. As was proved first by Rebsch (1938), the theoretical resolving power, due to spherical aberration, does not increase, but may decrease slightly with increasing electron speed. In a later paper, v. Ardenne (1941d) gives the following expression for the smallest resolved distance:

$$\delta = K \sqrt[4]{D_p} \sqrt[5]{V^3}, \qquad \ldots\ldots(45)$$

where D_p is the diameter of the bore of the objective pole-piece, K is a constant in which are lumped numerical coefficients and factors depending on the shape, saturation, degree of symmetry, etc., of the pole-pieces, and V is the accelerating voltage. These considerations, however, take into account only the resolving power of the front surface of the specimen, closest to the objective lens, and neglect to take into consideration the impaired resolving power of the layers of the specimen farther away from this lens. As illustrated in figure 31, in a thick specimen, due to scattering, the circle of least confusion necessarily increases with the distance from the objective lens. As the scattering decreases with increasing electron velocity, the resolving power for such thick specimens is, on the average, considerably improved, although a slight reduction of resolving power for the front layers might result from the use of such increased velocities.

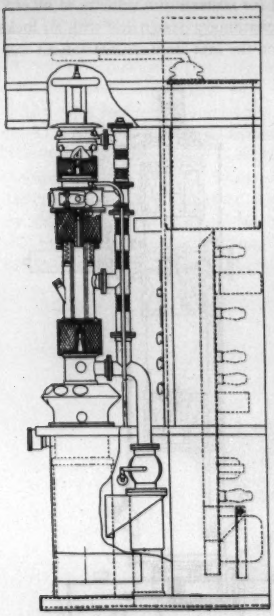


Figure 30. R.C.A. Type B electron microscope.

(By courtesy of Dr. V. K. Zworykin and Dr. J. Hillier.)

The next report on the construction of a higher-voltage instrument came from v. Ardenne (1941d). The instrument used for this purpose was the universal electron-microscope, described a year earlier, with an added electron gun to produce electron velocities up to 200 ekv. This gun was subdivided in three accelerating stages with each stage tied into the power supply. The paper gives a detailed study of the influence of optical aberrations on the image quality at higher velocities, and also of the influence of higher speed on the specimen. In particular, the question was raised to what extent living matter could be protected from destruction due to electron bombardment, and a device called an " object shadowing attachment " was incorporated for the observation of such specimens.

At the end of the same year, Zworykin, Hillier and Vance (1941 b) published a preliminary report on the development of a 300 kv. magnetic instrument. In this case also the microscope body was identical with the one reported earlier by the same investigators (Zworykin, Hillier and Vance, 1941a), with a two-stage

electron-gun added for the operation at high voltages. The power supply was built on the same principles as those described for the lower-voltage generator, i.e. the high voltage was produced by a high-frequency power supply, using a quadrupler rectifier system.

About the same time, M. Siegbahn constructed in Sweden an electron microscope of the transmission type with magnetic lenses. The maximum electron energy seems to be 80 ekv The instrument is of horizontal design and provided with means for stereoscopy. Up to the time of writing this *Report*, no extensive publication describing this instrument has come to the knowledge of the reviewer.

During the following year several papers appeared about modifications and additions to earlier instruments and also on new microscopes. In many ways one of these new instruments, reported by Bachman and Ramo (1942, 43 c), is

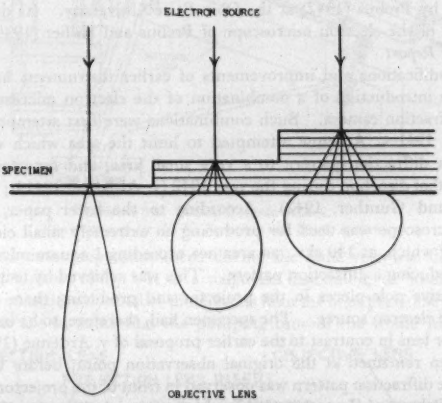


Figure 31. Schematic representation of the angular distribution of scattered electrons from different depths of a thick sample.

quite remarkable. The instrument, constructed in the research laboratory of the General Electric Company, is a two-stage electrostatic instrument. It does not have any air locks, and no provision is made for internal photography. By using a transparent fluorescent screen of very fine grain in end-on observation, the designers of the instrument deliberately renounced exploitation of the ultimate resolving power of such instruments. The guiding idea seemed to be to build a cheap and very much simplified instrument with a limited resolving power. In light optics there is a whole range of instruments, starting from the high-power research-microscope, down through the routine-testing instruments, to the inexpensive instruments for instruction. *A priori* it seems to be quite justified to follow the same principle in electron microscopy. The instrument, intended for ten times the resolving power of light microscopes, has not yet appeared on

the market as a commercial instrument. Its design is very compact, with both diffusion and mechanical pumps and all power supplies built in the base cabinet.

Another paper dealing with electrostatic instruments was published by Boersch (1942 b). This instrument, built at the University of Vienna, Austria, seems to be a very versatile microscope in allowing many different lens combinations for varying illumination and varying types of microscope observation. In particular, it allows a very simple transformation of the instrument from the conventional transmission type into an electron-shadow microscope, or into an electron-diffraction camera. The paper lists more than ten different methods for which the instrument can be used. This is the electron microscope which was used for the extremely interesting observations of Fresnel diffraction phenomena quoted earlier in this *Report*.

The third new instrument reported during the year 1942 is a magnetic microscope built by Prebus (1942) at the Ohio State University. Its design follows closely that of the electron microscope of Prebus and Hillier (1939) mentioned in an earlier *Report*.

The modifications and improvements of earlier instruments had mostly to do with the introduction of a combination of the electron microscope with an electron-diffraction camera. Such combinations were first attempted by Ruska (1940). In 1941 v. Ardenne attempted to limit the area which was used for producing a diffraction pattern to a very small area, and reported on further progress in that direction during the year 1942 (v. Ardenne, 1941 b; v. Ardenne, Schiebold and Gunther, 1942). According to the latter paper, the 200-kv. electron-microscope was used for producing an extremely small electron probe by means of which, at 140 ekv., an area not exceeding 1 square micron could be used for producing a diffraction pattern. This was achieved by using a reversed set of objective pole-pieces in the projector and producing there a very small image of the electron source. The specimen had, therefore, to be inserted below the projector lens in contrast to the earlier proposal of v. Ardenne (1941 b) where the specimen remained at the original observation point, before the objective lens, and the diffraction pattern was observed in front of the projector lens.

Hillier, Baker and Zworykin (1942) added a special diffraction adaptor to the R.C.A. electron microscope. This adaptor, replacing the projector lens, is operated by transferring the specimen from the object chamber into the adaptor and recording the diffraction pattern on the same photographic emulsion as the electron-microscope images.

v. Ardenne (1942) reported also on further development work on his universal electron microscope, One of the interesting details reported in this paper is the use of shortened object-cartridges for the purpose of producing different ranges of magnifications with the same instrument. Also reported is a new construction of magnetic pole-pieces for the objective lens, claiming a focal length of 0·9 mm. at 70-ekv. electron energy.

Two new instruments were announced during the year 1943. One instrument designed by Mahl and Pendzich (1943) is an entirely new approach to electron microscopy; it is an electron-mirror microscope, built with electrostatic lenses. The electron gun, condenser lens and objective lens are of similar design, as on earlier instruments. The projector lens, however, is used as an electron mirror

and produces the highly magnified image at the back surface of the objective lens. Obviously, the centre of the image is lost where an aperture has to be provided for the incident beam. The authors, however, produced relatively good images with this system. At present it is hard to judge whether it will have any further practical application beyond the novelty of the experiment.

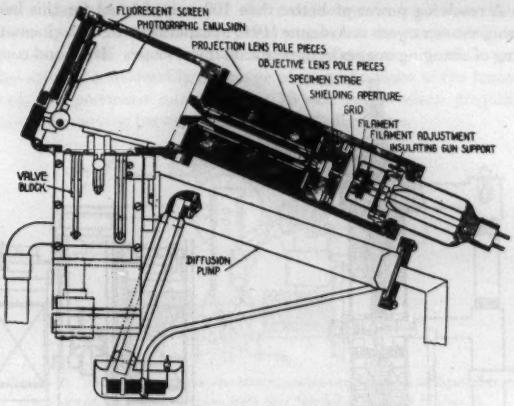


Figure 32. R.C.A. Console Model.
(*By courtesy of Dr. V. K. Zworykin and Dr. J. Hillier.*)

The second microscope announced during the same year was described by Zworykin and Hillier (1943). It is a magnetic instrument of 30-ekv. electron energy and of considerably smaller dimensions than the ones previously described. Its dimensions are so small that the pumping time of the whole instrument is

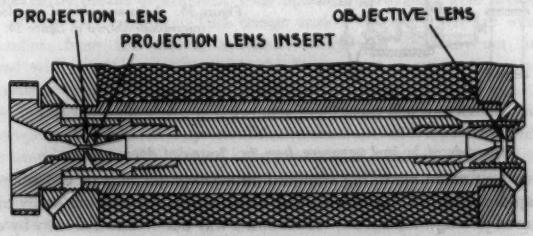


Figure 33. Cross-section of the pole-piece unit of the R.C.A. Console Model.
(*By courtesy of Dr. V. K. Zworykin and Dr. J. Hillier.*)

reduced to about 1 minute. This reduced volume is achieved by omitting completely the condenser lens and all the air locks, and merging the objective and projector lenses in a single unit. Figure 32 shows a cross-section of this instrument, with figure 33 giving details of the combined objective projector pole-piece

design. Due to this construction, the magnification range of the microscope becomes necessarily limited. This microscope has been, since its announcement, commercially produced under the name " Console Model ". The commercial instruments are built up at a slight angle from horizontal on top of a desk. The lower parts of the desk contain all the power supplies and the complete vacuum-plant. A resolving power of better than 100 A. is claimed for this instrument.

During the same year v. Ardenne (1943 a) experimented with cinematographic recording of changing images in the electron microscope. He found considerable

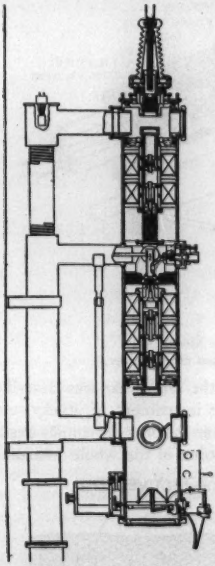


Figure 34. Electron microscope by Marton (1945 a).

(*Reproduced, by kind permission, from the 'Journal of Applied Physics'.*)

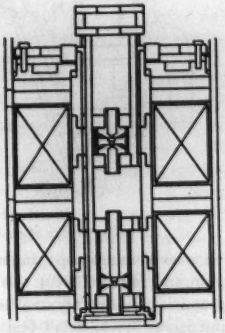


Figure 35. Lens assembly of the electron microscope shown in figure 34.

trouble at first when operating standard cinematograph film in vacuum. Due to the rapid drying out, the film became extremely brittle and was easily torn by the required jerking movement. He found a way out of this difficulty by operating the film-moving mechanism at a pressure of approximately 10^{-2} mm. Hg and producing the required pressure gradient by very intensive pumping of the adjacent parts of the microscope.

In the following year, 1944, the description of one new instrument was given by Marton (1944 c, 45 a). Figure 34 shows a cross-section of the microscope, and, some of its details are contained on figures 35 and 36. This instrument is of a magnetic type and is built for a maximum electron energy of 100 ekv. It is pro-

vided with air locks, both for specimen and photographic plates. Its mechanical stage is considerably modified, as compared with earlier designs, to provide a practically complete freedom from jerkiness. Improved facilities for stereoscopic microscopy, by means of a stage-tilting mechanism, have been used. This latter, by means of tilting angles up to $\pm 15\frac{1}{2}°$, extended the range of stereoscopic observation. The optical system of the microscope is built for three-stage magnification, instead of the customary two stages, for the purpose of including a much wider field of observation at low magnifications. Such a three-lens system also extends considerably the range of magnifications of the instrument; for instance, this particular microscope at 70 ekv. can produce magnifications in the instrument varying between 400 and 40,000 diameters.

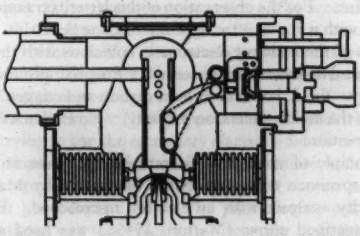


Figure 36. Object chamber of the electron microscope shown in figure 34.
(*Reproduced, by kind permission, from the 'Journal of Applied Physics'.*)

Another distinctive feature of this microscope is that the alignment of its optical elements is done by moving the pole-pieces in a plane perpendicular to the optical axis instead of moving the entire coils. This arrangement is shown on figure 35. The pole-pieces, held together by a brass sleeve, slide with very little clearance between corresponding iron disks. The lateral movement can be achieved by means of eccentric rods operated through a gear arrangement from outside the vacuum. This design makes the alignment of every lens independent from the others, a feature which can generally not be easily achieved with instruments in which the lower parts serve as support of the upper ones.

The last instrument which will be described briefly here is the R.C.A. type-U microscope. A detailed description of this instrument has not yet been published. At the time of writing this *Report*, only commercial releases, giving scant detail, are available. The maximum electron energy has been reduced, as compared to the R.C.A. type-B microscope, from 60 to 50 ekv. The inner volume of the instrument has been considerably reduced and a pumping time of about 1 minute achieved for it. In this way air locks could be completely eliminated. Otherwise, the design of the instrument seems to have followed, to a considerable extent, the design of the earlier R.C.A. microscopes.

In the same section, with the construction and design of electron microscopes, belong the descriptions of some auxiliary instruments and equipment in use for electron microscopy. Brief mention was made earlier of the different electron-diffraction adaptors and combinations for the use of the microscope for both

purposes. A more significant development is the use of the characteristic velocity losses, mentioned in the previous section, for recognition of the chemical constitution of the specimen. A special instrument, called an electron micro-analyser, was built for that purpose by Hillier and Baker (1943, 44). It consists of an inverted electron microscope using two optical stages for producing a reduced image of the electron source. The electron probe, which can be as small as 200 A. in diameter, enters, after passing through the specimen, through a third lens in a homogeneous magnetic field. This field is perpendicular to the direction of the beam and produces a velocity spectrum at 180° deflection. Without the deflecting field, the instrument can be used as an electron-shadow microscope, and the same area which was analysed can be viewed on a fluorescent screen at the end of the instrument. For the observation of this latter, a transparent fluorescent screen is provided with a 5 × to 30 × light microscope behind it. In preliminary tests, energy losses of the incident electrons in collisions with the inner electrons of a number of elements have been observed. Surprisingly the discrete energy loss corresponding to the *K* levels of at least carbon and oxygen have been found to be dependent on the incident electron velocity. No explanation was attempted for this latter observation.

The same principle of using the discrete energy losses on x-ray levels for determination of specimen composition has been used by Marton (1944 b) in combining a velocity analyser with an electron microscope. For this purpose, the microscope described above (Marton, 1945 a) was used with a deflecting system attached to the bottom of the instrument. Figure 37* shows the experimental arrangement with the " end-on " fluorescent screen of the microscope replaced by a disk provided with a fine slit. Through this slit, a part of the magnified electron image can enter a deflecting chamber. The deflecting system consists of a single electro-magnet producing direction-focusing at 135° of deflection. Observation of the velocity spectrum was made on the fluorescent screen, inclined at about 7° to the electron beam. For a selection of the proper part of the image of which the velocity spectrum is to be observed, the top surface of the supporting disk and of the slit are coated with fluorescent material. At 70 ekv. incident electron energy the instrument is able to detect energy changes of about 3 ev. Although at high energy losses (of the order of magnitude of 2 ekv. or more) the available intensity becomes exceedingly low, due to chromatic aberration of the lenses, this arrangement, which might be called an electron microspectrograph, offers the advantage of more direct observation, of adaptability to existing instruments provided with " end-on " screens, and of the possibility of using it for the determination of the chromatic aberration of lenses.

Another interesting auxiliary to electron microscopes is provided by the development of high-speed microtomes. Standard microtomes are limited to sections of 1 μ thickness and above. The useful upper limit of thickness for electron microscopy is, however, 0·5 μ. Different attempts have been made, therefore, to find a way out of this difficulty. v. Ardenne (1939 b) proposed the use of standard microtomes for cutting wedge-shaped sections. This method, however, does not seem to give very good results. A much more promising way was opened up by O'Brien and McKinley (1943), who operated the cutting knife at speeds of about 120 ft./sec. According to unpublished reports, Fullam, by

* Facing p. 221.

operating rotating knives at 600 ft./sec. cutting-speed, improved the method and is able regularly to produce microtome sections of 0·1 μ thickness.

Although this *Report* is mainly devoted to the conventional transmission-type electron microscopes, some remarks have to be made about developments of other related instruments during the last five years. Among such instruments we shall treat here briefly the scanning electron microscope and the emission-type electron microscope.

The scanning electron microscope, originally developed by v. Ardenne (1938 b, 38 d), encountered considerable difficulties, due to the smallness of the scanning probe and due to lack of contrast. The necessary smallness of the probe involved extremely low probe-current, and the instrument thus operated very close to the limits of the signal-to-noise ratio. The second limitation, that of contrast, is due to the fact that "image formation" is based on secondary electron emission. Unfortunately the secondary-to-primary ratio becomes practically independent of the nature of the emitting material for the high electron velocities used in electron microscopy.

Zworykin, Hillier and Snyder (1942) approached the problem slightly differently. Instead of trying to use the secondary electrons for direct recording of the scanned intensity, these secondaries were used to excite a fluorescent screen toward which they are accelerated. The fluorescent screen is placed in front of a multiplier-type photoelectric cell. The photoelectric cell is connected in more or less conventional way to a "facsimile" recorder. In this way a considerable reduction of "noise" is achieved. To obtain a higher contrast, the incident electrons are made to strike the specimen, at relatively low speed, where the secondary-to-primary ratio is more favourable. A resolving power of 500 A. has been claimed with this instrument.

Significant advances have been achieved also in emission microscopy. The problem of improvement of resolution of these instruments was attacked by Boersch (1942 a, 42 d), Brüche (1942), Dosse and Müller (1942), Mecklenburg (1942) and theoretically by Recknagel (1943). The most important result of these investigations was the demonstration that the resolving power primarily depends on the ratio $\frac{\epsilon}{V}$, where ϵ is the energy of the emitted electrons at the cathode surface, expressed in ev., and V is the accelerating voltage near the cathode. According to these calculations, the earlier attempts at emission microscopy necessarily failed because the field strength near the cathode was always much too low. By using specially shaped "immersion lenses", it can be demonstrated that high-resolution emission microscopy is possible. The best results up to now have been achieved by Mecklenburg, who showed resolutions below 400 A.

§ 5. TECHNIQUES AND APPLICATIONS OF ELECTRON MICROSCOPY

The techniques applied in electron microscopy may be divided into two main groups: (i) the techniques used for preparing the specimens, and (ii) the methods applied when using the electron microscope. As far as preparing the specimen is concerned, the most important considerations are convenience of operation and a reduction of ambiguity in the interpretation of the results. For

such purposes the specimen holders have been improved during the last years, the methods of preparing supports were to some extent changed, and new techniques were developed for the observation of specimens which were difficult or impossible to observe with earlier techniques.

The earliest design of electron-microscope specimen holders was a flat disk in which a small hole (about 0·1 mm. in diameter) was drilled (Ruska, 1934; Marton, 1934). Such a specimen holder had the disadvantage of being expensive and of offering difficulty in placing the specimen at the proper spot. It was replaced later (Marton, 1940) by flat disks cut out from fine wire-mesh. Woven wire-meshes were originally used of about 200 meshes per inch, and such meshes, or their electrolytic counterparts, have been adopted for all commercial instruments produced in the United States. As far as this reviewer can judge, the German instruments are still using the older type specimen mounts. When constructing the instrument last described, Marton (1945 a) introduced another type of specimen holder. It consists of wire mesh which is first rolled flat and then given a cup-shape, as indicated on figure 38. A plane surface of about 1 mm.

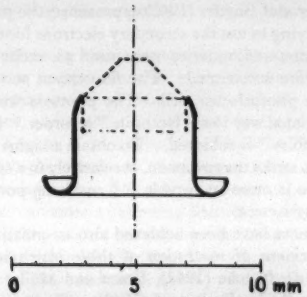


Figure 38. New type specimen holder.
(*Reproduced, by kind permission, from the 'Journal of Applied Physics'.*)

in diameter, and perpendicular to the optical axis, constitutes the observed part. It is joined by a conical surface of 45° to a cylindrical part fitting in a commercial eyelet. Such specimen holders have the following advantages: (i) they are easy to handle; (ii) specimens can be easily preparedat a well-defined part of the specimen holder; (iii) they are "polarized", i.e. the surfaces which support the collodion film or other films cannot be confounded with the "wrong" side of the specimen holder; (iv) they can be easily filed away for further use; (v) by means of a special tong, their insertion into, and withdrawal from, the microscope needs no particular skill on the part of the operator.

They can be easily adapted to commercial instruments, for instance by replacing the standard "cartridge" of the R.C.A. microscope with the "cartridge" shown on figure 39 (Marton, 1945 b).

The support in use for most electron-microscope specimens is still a thin film of collodion (Marton, 1936 a; v. Borries, Ruska, E. and Ruska, H., 1938; Ruska, H., 1939). Although far from being the best possible support, it has the great advantage of simplicity of preparation. The standard methods for making

such films are as follows: one or two drops of a 1½–2% solution of collodion in amyl acetate is allowed to spread out on a water surface of about 10 to 20 cm. diameter. After evaporation of the amyl acetate, the thin film is placed on the mounts previously described, either by immersing these holders previously in the water and lifting them out through the surface bearing the collodion film, or by placing the mounts, after formation of the film, on its upper surface and lifting it out with a special tool. The disadvantages of collodion film are that it is a rather fragile material and it tears away very easily under intense electron beams. This tearing is marked if the specimen placed on it consists of good insulators. The individual particles charge up and disrupt the film through the forces acting on the particles. A possible way out of this difficulty would be to use an oxidized collodion film, which has a much larger conductivity than the one normally used, as suggested by Abrams.*

A considerably better material for specimen supports is a plastic called "Formvar".† For equal thickness, it shows considerably more strength than

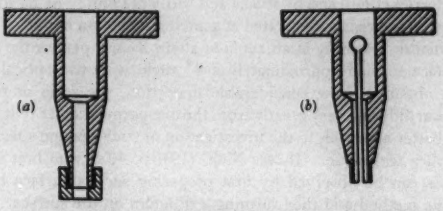


Figure 39. Adaptation of the specimen holder shown in figure 38 to R.C.A. instruments: (*a*) "Cartridge" of Type B microscope, (*b*) modified "cartridge". (*Reproduced, by kind permission, from the 'Journal of Applied Physics'.*)

collodion. It is, however, used much less frequently because of its more complicated preparation. Other supports showing excellent properties are aluminium-oxide films (Hass and Kehler, 1941) and evaporated silica.

The many ways of mounting specimens have necessarily to be adapted to the material which is investigated. In the case of biological specimens, it is important to avoid contamination by foreign biological material and impurities resulting from the drying-out of media used for preparing the material. In this respect, it is good to emphasize that only too often small residual crystals from saline or other material can be included, and modify considerably the appearance of the specimen. These difficulties are somewhat reduced when using the microscope for investigation of chemical substances, but they are by no means excluded. For specific details of such techniques, we refer to articles by Marton (1941), Anderson (1942), Marton (1943), Mudd and Anderson (1944).

An extremely interesting new technique for preparing specimens was developed by Williams and Wyckoff (1945 a, 45 c). Originally developed for the thickness or height determination of small objects (Williams and Wyckoff, 1944), it consists of evaporating in vacuum a thin layer of gold or chromium (or in some

* Private communication.

† Manufactured by the Shawinigan Resins Corp., Empire State Building, New York.

cases uranium) on to the surface of the prepared specimen, whereby the atomic beam makes a small grazing angle with the surface. The resulting micrographs are printed in negative to show the shadows as black on white instead of the less conventional white on black of the positive prints. The pictures thus obtained are extremely plastic and allow the recognition of considerably smaller particles than hitherto observed. So, for instance, this technique reveals for the first time the proper structure of the collodion films, which had not previously been observed. As an illustration of this technique, figure 40 shows "gold-shadowed" particles of tobacco mosaic virus.

A very important field for which no observation techniques existed five years ago, is the investigation of non-transparent solids. By non-transparent solids are meant all the specimens which cannot be prepared in layers sufficiently thin to be observed directly in the electron microscope. All the problems of metallography belong in this class, with many examples taken from mineralogy, palæontology and other fields. A first attempt to observe such specimens was made by v. Borries (1940) and by Ruska and Müller (1940). As no true reflection exists in electron optics, they operated at glancing reflection by tilting the electron gun and condenser assembly at an angle of about 8° and placing the specimen so that the surface formed approximately a 4° angle with the optical axis. The images thus obtained show considerable distortion, the scale of reproduction in one direction differing very greatly from the one perpendicular to it.

A much better approach to the investigation of such opaque specimens is the so-called replica technique. It was Mahl (1940b, 40c) who first showed that metal surfaces can be observed by first preparing such a surface by the usual metallographic method and then forming a thin film on the surface. The films, after being spearated from the base surface, can be used for transmission observation in the electron microscope. If properly prepared, they show a great wealth of detail and exceed considerably the limits set by the resolving power of light microscopes.

The first replica technique of Mahl consisted in producing a thin structureless film, generally of oxide, on the previously prepared surface, and then removing the film by chemical means. The oxidation can be produced by chemical, thermal or electrolytic methods. This method has the disadvantages that it applies only to metals which form structureless surface films, like aluminium, aluminium alloys, nickel, iron, and a nickel-beryllium alloy, and that to remove the film the base metal has to be, at least partially, destroyed.

Mahl also described a replica technique carried out by using thin lacquer films on the surface and dissolving the base metal. This latter method was developed into techniques mostly in use at present, methods avoiding the destruction of the base metal. The latest methods can be divided roughly in two groups: the single-step method, producing a "negative" replica, and the two-step method, producing a "positive" replica.

The negative replica method uses plastics or other substances which can be stripped directly from the surfaces without modifying it. As an example, the so-called Formwar technique (Schaefer and Harker, 1942; Schaefer, 1943; Harker and Murphy, 1945) seems to be most representative, although other materials, like collodion (Thomassen, Williams and Wyckoff, 1945) or gelatine

Figure 40. Tobacco mosaic virus with uranium "shadowing". Note a bacterial cell with its long flagella and the relatively rough background.
(Specimen by courtesy of Prof. T. E. Rawlins, University of California.)

To face page 240

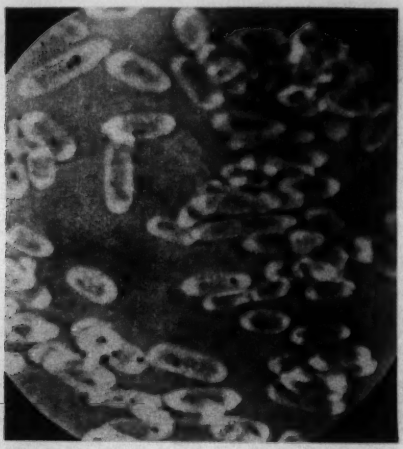


Figure 41. Unusual case of specimen modification through electron bombardment. Practically the whole cell content of *Xanthomonas begoniae* " evaporated " during observation, whereas the carbohydrate background in which the bacterial cells are imbedded remained unchanged.

(*Specimen by courtesy of Dr. M. P. Starr, Hopkins Marine Station.*)

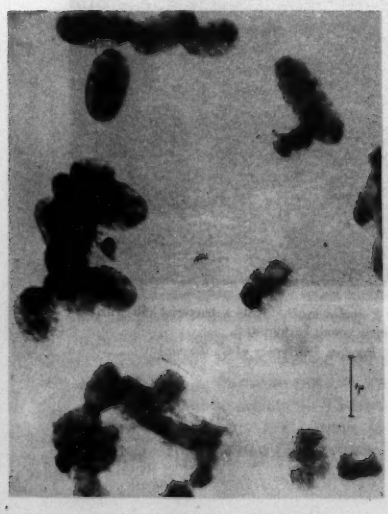


Figure 43. *B. pestis* treated with formaldehyde.
(*Specimen by courtesy of Prof. K. F. Meyer, University of California.*)

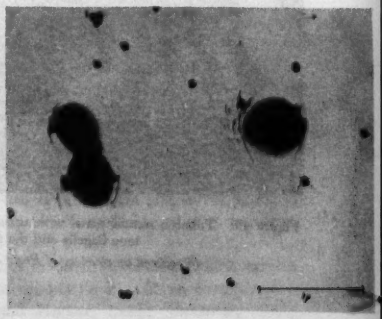


Figure 44. *Pyocyaneus* bacteriophage.
(*Specimen by courtesy of Prof. E. W. Schultz, Stanford University.*)

(Marton and Schiff, 1941), have been stripped from metal surfaces without the destruction of the base. The two-step method involves first making a relatively thick negative replica, which can be separated from the base without trouble and on which a thin positive replica is made. The thick negative replica can then be dissolved. A first attempt to use this two-step method used silver as the material for the first negative replica and collodion as material for the positive replica (Zworykin and Ramberg, 1941). Considerably better success was achieved by a method developed by Heidenreich and Peck (1942) consisting of moulding polystyrene, at relatively high pressure (and moderate temperatures), on to the metal surface, breaking off the polystyrene mould and evaporating, in vacuum, vitreous silica on to the polystyrene surface. The actual techniques are well described in detail in a number of publications from which Barrett (1943), Zworykin (1943), Keller and Geisler (1944 b) and Harker and Murphy (1945) may be cited.

Many specimens used in electron microscopy contain water or have water as their normal medium of existence. It has been emphasized since the beginning of electron microscopy that water losses, due to the specimen being in a very good vacuum, may modify considerably the observed structures. The different attempts to avoid such water losses up to now have not been very successful. Following some experiments by Marton in 1935, and a patent description by v. Borries and Ruska, H. (1941), Abrams and McBain (1944) attempted to build electron-transparent and closed cells for the observation of specimens in their natural media. Although the cells proved to be vacuum tight, electron-optical observation through them up to now have completely failed. The use of much higher electron speeds may offer better chances for this technique.

A better way possibly to obtain observations of unmodified structures, or structures with a minimum of modification, is to freeze them and then dehydrate by keeping them at low temperature. Such techniques have been applied repeatedly in different fields of biological research and seem to be successful. In electron microscopy these methods have not been introduced sufficiently to permit a definite opinion about their value.

Once the specimen is properly prepared, it is important that it be handled correctly in the electron microscope. Such correct technique enables the most advantageous use to be made of the electron microscope for the observation of the specimen. The conditions for best observation are: (i) choice of the right electron velocity; (ii) choice of the right beam intensity; (iii) correct alignment of the instrument; (iv) choice of the illuminating aperture; (v) choice of the objective aperture; and (vi) selection of the best-suited photographic material.

The electron velocity selected for observation of a particular specimen depends entirely on the properties of the specimen. In the case of very small organic particles, like virus or other high molecular organic compounds, relatively low velocities can be used. Such low velocities would insure *a priori* the highest possible contrast. This viewpoint, however, has to be slightly modified since Williams and Wyckoff introduced their gold-shadowing technique. With this technique even virus particles or protein molecules can be most advantageously observed at high electron velocities. When it comes to observing thicker specimens, like, for instance, bacterial cells, internal structure sometimes is not revealed until

beams of relatively high velocity are used. For the average specimen, electron energies ranging between 60 to 100 ekv. seem to offer the best range.

To avoid the modification of specimens or the formation of artefacts, it is important to operate at sufficiently low electron-beam intensity. It has been observed repeatedly that some specimens are more sensitive than others and that, after prolonged observation, the appearance of some specimens may considerably change (v. Borries and Ruska, E., 1939 b; Lembke, Ruska, H., and Christophersen 1940; Schmidt, 1943). Such changes are due mostly to the temperature rise of the specimen. v. Borries and Glaser (1944) made an attempt to calculate, to a first approximation, the temperature rise in an electron-microscopical object during observation. These calculations give estimated temperatures which, for instance for a bacterial cell supported by a collodion film, might be as high as 300° c. In the same paper they present experimental evidence for the statement that bacteria like, for instance, those of diphtheria do not change in appearance after having been heated to 300° c. Similar observations have been reported on other bacterial cells by Mudd and Anderson (1944). It has to be pointed out, however, that experience in this laboratory shows that, although most bacterial cells support very well an intense electron bombardment, some will change markedly after only short exposure to the electron beam, as shown in figure 41. A correlation between bacterial structure and the ability to change appearance rapidly is yet completely missing. Visual observation indicates that some components may evaporate under the intense electron beam while others do not. To avoid such changes, Marton (1936 a) proposed a blind-focusing method on a " dummy " specimen, and this proposal has been worked out in more detail by v. Ardenne (1941 d) in his proposed object-shadowing attachment. This latter consists of a wedge-shaped " shutter " intercepting part of the illuminating beam and thus protecting part of the specimen from irradiation during the visual focusing of the image. After the best focus has been achieved, and the intensity reduced to the absolute minimum necessary for photography, the shutter is withdrawn. Thus photography of the specimen without prior irradiation is made possible.

For achievement of the best results, it is important that the different optical elements of the instrument should be properly aligned. Extensive publications on this subject are curiously missing. The instruction manuals of the R.C.A. commercial instruments, which are available only to users of those instruments, contain information about the necessary procedure for obtaining correct alignment. As the optical axis of the instrument is generally a curved one, the criterion for the best alignment is judged from observation of the rotation of the successive images around themselves, when the lens strength is varied. The best procedure, therefore, consists in first using the instrument without all its lenses and successively aligning each optical element (including the electron source) by switching on the missing lenses as the alignment proceeds. The actual process might vary, depending on which point of the optical system is the fixed point.

v. Borries and Ruska, E. (1939 c) discussed in considerable detail the best conditions for the aperture of the illuminating beam. Although their conclusions have been since modified by Marton and Hutter (1944 a, 44 b), the best procedure for obtaining the necessary low illuminating aperture consists, as they indicated,

of increasing the condenser lens-strength beyond the strength necessary for producing an image of the source on the specimen. Under these conditions, the actual illuminating aperture is the angle subtended from the (real) image of the source formed between the specimen and the condenser lens. Illuminating apertures of the order 10^{-5} radian can be produced this way. In a recent paper, Hillier and Baker (1945) discuss the conditions of image formation with such illuminating apertures and come to the conclusion that only too often variations in the quality of the micrographs result from inaccuracies in the adjustment of the illuminating system. They propose, therefore, to replace the reduction of the illuminating aperture done earlier by changing the condenser lens-strength by a different procedure, consisting of keeping the condenser lens-strength constant and introducing a limiting aperture of required size in the condenser lens. In practice such a proposal is reduced to providing the condenser lens with exchangeable apertures, one for focusing and one for photographic purposes, and resembles the procedure used earlier by v. Ardenne (1940a).

A considerable amount of discussion has been devoted to the size of the objective aperture. Whereas it has been demonstrated theoretically (v. Ardenne, 1938a; Rebsch, 1938), that the best objective aperture is of the order of magnitude of 10^{-3} radian, this by no means indicates that a physical aperture of that size is needed in the objective lens. v. Borries and Ruska (1939c) demonstrated that, for thin objects at least, the image quality is primarily defined by the solid angle of the incident (condenser) beam alone. A survey of many microscopes operated in the United States seems to indicate that about 50% of the users of the microscope prefer to have a limiting aperture, whereas the other 50% prefer not to have any limiting aperture in the objective lens at all. Scattering considerations show that the correct answer is that the need for an aperture depends entirely on the nature of the specimen. Generally speaking it is advantageous to use a limiting aperture with specimens of low density giving a predominance of small scattering angles, whereas for specimens giving large-angle scattering, the pole-pieces alone might provide enough physical limitation of the scattered beam. It is also important that the limiting aperture be kept free of impurities because such impurities, by accumulating electrical charges, might considerably modify the steadiness of the image.

A fair number of papers have been devoted to the subject of the recording of the electron-microscope images. The information on the electron sensitivity of different photographic material, available before the use of the electron microscope, is rather scarce. Obviously the best photographic material for recording an electron micrograph has to satisfy the following requirements: (i) high sensitivity to electron irradiation; (ii) small grain-size to insure the best possible resolving power; and (iii) a convenient gradation for recording on a single exposure all the details of the image.

A first, more or less qualitative, comparison of photographic material for use in electron microscopy was published by Marton (1939b). Both the lantern-slide and process-type emulsions seemed to offer advantages for recording, with the process emulsion showing slightly higher sensitivity.

During the same year, v. Ardenne (1939a), considering the adaptability of photographic emulsions for electron-microscope investigation, came to the

conclusion that Schumann-type emulsions are most desirable. A year later, in his book, v. Ardenne (1940 a) published more detail about his investigations into the fitness of such emulsions, and his results seemed to indicate a definite advantage in the use of Schumann plates. After this a number of papers on this subject were published by Baker, Ramberg and Hillier (1942), v. Borries (1942 a–44), v. Ardenne (1943b) and Nissen (1944).

In a number of papers v. Borries reported on his investigation of the blackening of photographic emulsions by electrons and their correlation with the intensity distribution in the electron-microscopic image. For this purpose, quantitative measurements were carried out on the blackening and the corresponding current-density per unit surface of the emulsion. The photographic material investigated by v. Borries was entirely of German origin. However, some later measurements allow us to compare his findings with other photographic material. For quantitative measurements, he introduces first the notion of a " work density " (*Arbeitsdichte*) defined as the amount of energy per unit surface needed to produce a certain standard blackening. As such a standard, the photographic density 0·5 was adopted by him, the " work density " for that blackening being given by

$$a_{0\cdot5} = V q_{0\cdot5}, \qquad \ldots\ldots(46)$$

where V is the electron energy in ev. and $q_{0\cdot5}$ is the amount of charge, measured in coulomb per cm^2., required for producing a blackening $D = 0\cdot5$. For estimating the fitness of an emulsion, he introduced the notion of *Bildpunktarbeit*,* defined as

$$A = a_{0\cdot5}\delta_p^{\,2}, \qquad \ldots\ldots(47)$$

whereby δ_p is the least resolved distance on the photographic emulsion. This *Bildpunktarbeit* plotted against the electron energy is reproduced in figure 42 for two different contrasts. It is a marked function of the electron velocity, showing a minimum for most emulsions in the region of about 50 ekv. The resolving power of the emulsion itself depends both on the graininess and on the available contrast. v. Borries (1942 a) also carried out measurements on the total amount of silver (salt) contained in the emulsion per unit surface and found, as could be expected, a direct relationship between the optimum electron velocity and the weight of the emulsion.

Obviously another variable has to be investigated, and that is the rôle of the developer in determining the quality of the image. An extensive investigation of this factor was contained in v. Borries' last paper (1944) where also an estimate of the minimum of *Bildpunktarbeit* required for recording a living specimen was computed. To summarize his results, he came to the conclusion that the best available emulsions still require 3×10^{-12} watts per image element which according to his estimates is beyond the lethal dosage for most biological specimens. In sharp contrast to v. Ardenne's results, he finds the Schumann emulsion rather insensitive, provoking a controversy about this subject (see v. Ardenne, 1943 b; v. Borries, 1944; Nissen, 1944). A further interesting point in v. Borries' investigations (1942 c) is the attempt to use photographic emulsions with absorbing layers on top of them for slowing down very high-speed electrons to the optimum

* This reviewer tried vainly to translate this expression into what he estimated to be correct English. Literally " image point work " would express it perhaps, but it does not seem to be very satisfactory.

velocity for the photographic emulsions. This method was also investigated during the same year by Baker, Ramberg and Hillier (1942). In their paper, besides giving the results of quantitative measurements for the Eastman medium

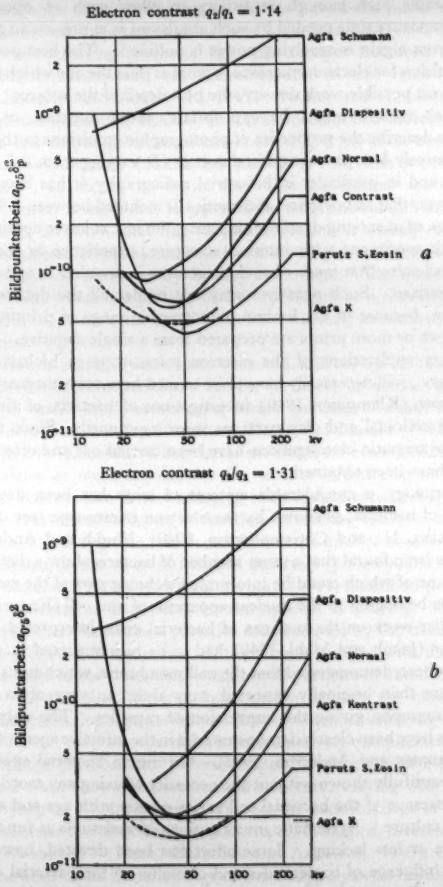


Figure 42. "Bildpunktarbeit" vs. electron energy for two different contrasts (v. Borries, 1942 c).

lantern-slide plates, they also describe experiments in which the emulsion was covered with a retarding layer. Such plates were found to give an advantage for high electron velocities.

The attempts to use extremely fine-grain photographic emulsions and to reduce exposure times by operating at lower magnifications in the electron microscope itself, seem to have failed. None of the emulsions of extreme resolving power possesses high enough sensitivity to allow such an operation. The increase of exposure time needed by such emulsions is at present so great that the advantage from a gain in resolving power is nullified. The best possible photographic emulsion for electron-microscope work is thus the one which will produce, with the lowest possible work density, the best detail of the subject.

A further remark might be appropriate about contrast considerations. Most papers describe the properties of photographic emulsions in the neighbourhood of relatively low photographic densities (D varying from 0·5 to 1·0). In x-ray work, and in particular in industrial radiography, it has been recognized lately, however, that an operation at densities D included between 2·0 to 3·0 offers the advantage of observing details which are quite lost at lower densities. In the case of thick specimens with internal structure, experience in this reviewer's laboratory indicates that more of such structures is revealed in apparently over-exposed negatives. Such negatives may not render all the details on a single positive print, because of the limited photographic range of printing papers, in which case two or more prints are prepared from a single negative.

The many applications of the electron microscope in biology, chemistry, metallurgy, etc., will necessarily have to be treated here very summarily. Earlier in these *Reports* (Klemperer, 1940), investigations of bacteria, of virus particles, and of some colloidal and clay particles were mentioned. Since that date, in many fields systematic investigations have been carried out and often suprisingly new results have been obtained.

In bacteriology, a considerable amount of work has been devoted to the morphology of bacteria, as shown by the electron microscope (see, for instance, Lembke, Ruska, H., and Christopherson, 1940; Mudd and Anderson, 1944, etc.). It has been found that a great number of bacteria show a distinct internal structure, some of which could be interpreted as being part of the cytoplasm and other parts as belonging to the nuclear apparatus of the cell (Knaysi and Mudd, 1943). Earlier work on the outlines of bacterial cells, interpreted at that time as capsulation (Jakob and Mahl, 1940), had to be re-interpreted in almost every case. It has been demonstrated that the cell membrane, which has a much more rigid structure than originally expected, may show up very often in bacterial electron micrographs, giving the impression of capsules. The only case where such capsules have been clearly demonstrated is in the infective agent of pneumonia (Mudd, Heinmets and Anderson, 1943). On motile bacterial species, flagella have been beautifully shown without the necessity of using any mordant or stain.

The appearance of the bacterial cell varies greatly with age and treatment of the bacterial culture. Systematic investigations of variations as functions of age are still more or less lacking. Some effort has been devoted, however, to the study of the influence of bactericides and of stains on the bacterial cell (Mudd, 1943). Generally speaking, bactericides and stains produce a coagulation or shrinkage of the inner contents of the cell with increased contrast in many cases as shown on figure 43.* Adsorption reactions also have been studied on bacterial cells, with special attention to the relationship of such reactions to immunity

* Facing page 241.

(Mudd, 1944). Such reactions are better discussed later, in that section of the *Report* dealing with antibody-antigen reactions on viruses.

The microscopical study of viruses truly starts with the advent of the electron microscope. Although some of the largest viruses have been made visible by means of ultra-violet microscopes, most of these organisms have become observable only since the resolving power of light microscopes has been exceeded. The first systematic investigation was that of tobacco mosaic virus (Kausche, 1940 a, 40 b; Kausche and Ruska, H., 1940; Stanley and Anderson, 1941). Ultra-centrifuge and streaming double-refraction observations indicated earlier the existence of rod-shaped particles of about $430 \times 12{\cdot}3$ mμ dimensions. The electron microscope confirmed that the unit particle of this virus is a rod-shaped " macromolecule ", as shown on figure 40. The dimensions are somewhat different, the diameter being about 15 mμ and the length about 280 mμ. It has to be emphasized, however, that this length dimension is merely a statistical average with considerable spread and that some investigators (Frampton, 1942), do not agree with views expressed by Stanley and others. The question is still open to some extent whether a particle of 280 mμ length or a shorter one is the fundamental unit, or whether the small particles are merely breakdown products of the larger ones.

Many other virus particles have been observed, and several of these investigations have produced controversies. So, for instance, influenza virus, which was described in earlier work as having spherical particles of about 120 mμ in diameter, was described later by Chambers, Henle and Lauffer (1943) as being only about 10 mμ in diameter. These latter investigators considered for a while the larger particles as mere carriers for the smaller infective agent. It seems to this reviewer that opinion is now swinging back toward considering the larger particles as the true infective agents. Another interesting example of diversity of views on such infective agents, which are close to the limit of observation in the electron microscope, is the virus of infantile paralysis (poliomyelitis). Tiselius and Gard (1942), Gard (1943 a, 43 b) and Melnick (1944) described poliomyelitis virus as long filamentous particles. In the opinion of Loring, Marton and Schwerdt (1944), such filamentous particles are produced by the chemical treatment during purification of the virus. If the virus is purified with the exclusion of the chemical part of the treatment, small round particles of about 15 to 20 mμ are observed. In the class of virus investigation belongs the observation of bacterial viruses (bacteriophage particles). One of the most suprising facts about investigation of these particular particles is their relatively complicated structure. Most bacteriophage particles are " sperm-shaped ", consisting of a head and a tail. The head often shows distinct internal structure, varying with the nature of the bacteriophage. So, for instance, whereas the bacteriophage of *Escherichia coli* shows head structures described as " X-shaped, Z-shaped, inverted Z-shaped and diplococcus-shaped " (Luria, Delbruck and Anderson, 1943), the internal structure of the head of the bacteriophage of *Bacillus pyocyaneus* has a filamentous internal structure resembling a somewhat distorted figure 8, as shown by figure 44 * (Schultz, Thomassen and Marton, 1944).

Experimental investigation of adsorption reactions on viruses demonstrated that the earlier antibody-antigen reaction theories of Pauling described the

* Facing page 241.

observed phenomena rather closely. As has been shown by Anderson and Stanley (1941), the adsorption reaction consists of a more or less uniform covering of the virus particle by the adsorbed material.

Besides purely microbiological observations, the electron microscope has also been used for the investigation of relatively large objects, like the structure of butterfly wings, of insect trachea, etc. (Anderson and Richards, 1942; Richards and Anderson, 1942).

The number of chemical investigations carried out by means of the electron microscope is perhaps even greater than the biological ones. A brief enumeration of chemical substances and chemical reactions includes: Smoke particles, (Friess and Müller, 1939; Barnes and Burton, 1941), clays, (Eitel and Schusterius, 1940; Shaw, 1942), and other ceramic raw materials (Eitel, 1943 a, 43 b), cement and lime (Clark, Bernays and Tordella, 1942; v. Ardenne and Endell, 1942; Sliepcevich, Gildart and Katz, 1943), colloidal carbon, (Wiegand and Ladd, 1942; Hofmann, 1943; Mrgudich and Clock, 1944), and its rôle in rubber reinforcement (Schoon and Koch, 1941; Ladd, 1944), the study of rubber and plastics (Hall, Hauser, LeBeau, Schmitt and Talalay, 1944), cellulose (Barnes and Burton, 1943; Franz, Schiebold and Weygand, 1943; Husemann, 1943), soaps (Marton, McBain and Vold, 1941), catalysts (Schoon and Beger, 1941; Turkevich, 1945), photographic emulsions (v. Ardenne, 1940 d; Hall and Schoen, 1941), etc.

In metallography, publications on the use of the electron microscope include investigations of the structures of many metals and alloys (Mahl, 1941; Mahl and Pawlek, 1942; Mahl and Stranski, 1943; Keller and Geisler, 1944 a, 44 b). Some recent investigations deal with specific problems, like age-hardening of alloys (Harker and Murphy, 1945), the formation of martensite in steel (Semmler-Alter, 1942), or the spacings of slip lines of precipitated particles and of pearlite lamellae (Barrett, 1943, 44).

It may also be briefly mentioned that the electron microscope is being applied to some extent to the solving of purely physical problems. As examples in that direction, the investigation of evaporated metal-films (Hass, 1942; Picard and Duffendack, 1943), of the active deposit from radiothorium, (Martin, Baylor and Clark, 1944), and of ferromagnetic domains (Marton, 1944 a) may be mentioned.

ACKNOWLEDGMENT

The author wishes to express his indebtedness to his friend, Professor Paul H. Kirkpatrick, for critical reading of the manuscript.

REFERENCES

Abrams, I. M. and McBain, J. W., 1944. *J. Appl. Phys.* **15**, 607.
A. E. G. Forsch.-Inst., Jahrbuch, 1940. *Sonderheft*: *Übermikroskop*, **7**, 1–90 (Berlin . Springer).
Anderson, T. F., 1942. *Advances in Colloid Science*, **1**, 353 (New York : Interscience).
Anderson, T. F. and Richards, A. G., 1942. *J. Appl. Phys.* **13**, 748.
Anderson, T. F. and Stanley, W. M., 1941. *J. Biol. Chem.* **139**, 339.
v. Ardenne, M., 1938 a. *Z. Phys.* **108**, 308.
v. Ardenne, M., 1938 b. *Z. Phys.* **109**, 553.
v. Ardenne, M., 1938 c. *Z. Phys.* **111**, 152.
v. Ardenne, M., 1938 d. *Z. tech. Phys.* **19**, 407.
v. Ardenne, M., 1939 a. *Z. Phys.* **114**, 379.

v. Ardenne, M., 1939 b. *Z. wiss. Mikr.* **56**, 8.
v. Ardenne, M., 1940 a. *Elektronen Übermikroskopie* (Berlin : Springer).
v. Ardenne, M., 1940 b. *Naturwissenschaften*, **28**, 113.
v. Ardenne, M., 1940 c. *Z. Phys.* **115**, 339.
v. Ardenne, M., 1940 d. *Z. angew. Photogr.* **2**, 14.
v. Ardenne, M., 1941 a. *Naturwissenschaften*, **29**, 780.
v. Ardenne, M., 1941 b. *Z. Phys.* **117**, 515.
v. Ardenne, M., 1941 c. *Z. Phys.* **117**, 602.
v. Ardenne, M., 1941 d. *Z. Phys.* **117**, 657.
v. Ardenne, M., 1942. *Phys. Z.* **43**, 11.
v. Ardenne, M., 1943 a. *Z. Phys.* **120**, 397.
v. Ardenne, M., 1943 b. *Z. Phys.* **121**, 1.
v. Ardenne, M., and Beischer, D., 1940. *Z. Elektrochem.* **46**, 270.
v. Ardenne, M. and Endell, K., 1942. *Zement*, 313.
v. Ardenne, M. and Friedrich-Freksa, H., 1941. *Naturwissenschaften*, **29**, 523.
v. Ardenne, M., Schiebold, E. and Günther, F., 1942. *Z. Phys.* **119**, 352.
Bachman, C. H. and Ramo, S., 1942. *Phys. Rev.* **62**, 494.
Bachman, C. H. and Ramo, S., 1943 a. *J. Appl. Phys.* **14**, 8.
Bachman, C. H. and Ramo, S., 1943 b. *J. Appl. Phys.* **14**, 69.
Bachman, C. H. and Ramo, S., 1943 c. *J. Appl. Phys.* **14**, 155.
Baker, R. F., Ramberg, E. G. and Hillier, J., 1942. *J. Appl. Phys.* **12**, 450.
Barnes, R. B. and Burton, C. J., 1941. *Amer. Chem. Soc., News Edn.*, **19**, 965.
Barnes, R. B. and Burton, C. J., 1943. *Industr. Engng. Chem., Industr. Edn.*, **35**, 120.
Barrett, C. S., 1943. *Trans. Amer. Inst. Min. Metall. Engrs., Tech. Publ.* No. 1637, and *Metals Tech.* **10**, No. 6, 1.
Barrett, C. S., 1944. *J. Appl. Phys.* **15**, 691.
Beischer, D., 1943. *Kolloidchem. Taschenb.*, 2 Aufl., 425.
Boersch, H., 1939. *Z. tech Phys.* **20**, 346.
Boersch, H., 1940. *Naturwissenschaften*, **28**, 709.
Boersch, H., 1942 a. *Naturwissenschaften*, **30**, 120.
Boersch, H., 1942 b. *Phys. Z.* **43**, 515.
Boersch, H., 1942 c. *Z. Phys.* **118**, 706.
Boersch, H., 1942 d. *Z. tech. Phys.* **23**, 129.
Boersch, H., 1943 a. *Phys. Z.* **44**, 32.
Boersch, H., 1943 b. *Phys. Z.* **44**, 202.
Boersch, H., 1944. *Kolloidzschr.* **106**, 169.
v. Borries, B., 1940. *Z. Phys.* **116**, 370.
v. Borries, B., 1942 a. *Phys. Z.* **43**, 190.
v. Borries, B., 1942 b. *Z. angew. Photogr.* **4**, 42.
v. Borries, B., 1942 c. *Z. Phys.* **119**, 498.
v. Borries, B., 1944. *Z. Phys.* **122**, 539.
v. Borries, B. and Glaser, W., 1944. *Kolloidzschr.* **106**, 123.
v. Borries, B. and Ruska, E., 1939 a. *Naturwissenschaften*, **27**, 281.
v. Borries, B. and Ruska, E., 1939 b. *Naturwissenschaften*, **27**, 577.
v. Borries, B. and Ruska, E., 1939 c. *Z. tech. Phys.* **20**, 225.
v. Borries, B. and Ruska, E., 1939 d. *Z. wiss. Mikr.* **56**, 317.
v. Borries, B. and Ruska, E., 1940. *Naturwissenschaften*, **28**, 366.
v. Borries, B. and Ruska, E., 1941. *Ergebn. exakt. Naturw.* **19**, 237.
v. Borries, B. and Ruska, H., 1941. *U.S. Patent* 2,256, 191.
v. Borries, B., Ruska, E. and Ruska, H., 1938. *Klin. Wschr.* **17**, 921.
Brüche, E., 1942. *Kolloidzschr.* **100**, 192.
Brüche, E. and Haagen, E., 1939. *Naturwissenschaften*, **27**, 809.
Brüche, E. and Recknagel, A., 1941. *Elektronengeräte, Principien und Systematik* (Berlin: Springer).
Busch, H., 1927. *Arch. Elektrotech.* **18**, 583.
Burton, E. F. and Kohl, W. H., 1942. *The Electron Microscope* (New York: Reinhold)
Chambers, L. A., Henle, W. and Lauffer, M. A., 1943. *J. Exp. Med.* **77**, 265.
Clark, G. L., Bernays, P. M. and Tordella, J.P., 1942. *Proc. 24th Natl. Lime Ass.*
Dosse, J., 1941 a. *Z. Phys.* **117**, 316.

Dosse, J., 1941 b. *Z. Phys.* **117**, 437.
Dosse, J., 1941 c. *Z. Phys.* **117**, 722.
Dosse, J. and Müller, H. O., 1942. *Z. Phys.* **119**, 415.
Eitel, W., 1943 a. *Ber. dtsch. keram. Ges.* **24**, 37.
Eitel, W., 1943 b. *Abh. preuss. Akad. Wiss.* 5.
Eitel, W. and Schusterius, C., 1940. *Chem. d. Erde*, **13**, 322.
Frampton, L. V., 1942. *Science*, **95**, 232.
Franz, E., Schiebold, E. and Weygand, C., 1943. *Naturwissenschaften*, **31**, 350.
Friess, H. and Müller, H. O., 1939. *Gasmaske*, **11**, 1.
Gard, S., 1943 a. *Klin. Wschr.* **22**, 315.
Gard, S., 1943 b. *Arch. ges. Virusfschg.* **3**, 1.
Glaser, W., 1940 a. *Z. Phys.* **116**, 19.
Glaser, W., 1940 b. *Z. Phys.* **116**, 56.
Glaser, W., 1941. *Z. Phys.* **117**, 285.
Glaser, W., 1943. *Z. Phys.* **121**, 647.
Glaser, W. and Lammel, E., 1941. *Ann. Phys., Lpz.*, **40**, 367.
Hall, C. E., Hauser, E. A., LeBeau, D. S., Schmitt, F. O. and Talalay, P., 1944. *Industr. Engng. Chem., Industr. Edn.*, **36**, 634.
Hall, C. E. and Schoen, A. L., 1941. *J. Opt. Soc. Amer.* **31**, 281.
Harker, D. and Murphy, M. J., 1945. *Trans. Amer. Inst. Min. Engrs., Tech. Publ.* No. 1811.
Hass, G., 1942. *Kolloidzschr.* **100**, 230.
Hass, G. and Kehler, H., 1941. *Kolloidzschr.* **95**, 26.
Hawley, G. G., 1945. *Seeing the Invisible* (New York : Knopf).
Heidenreich, R. D., 1942. *Phys. Rev.* **62**, 291.
Heidenreich, R. D. and Peck, V. G., 1942. *Phys. Rev.* **62**, 291.
Heidenreich, R. D. and Sturkey, L., 1945. *J. Appl. Phys.* **16**, 97.
Hillier, J., 1939. *Canad. J. Res.* **17**, 64.
Hillier, J., 1940. *Phys. Rev.* **58**, 842.
Hillier, J., 1941. *Phys. Rev.* **60**, 743.
Hillier, J., 1943. *Phys. Rev.* **64**, 318.
Hillier, J. and Baker, R. F., 1943. *Phys. Rev.* **61**, 722.
Hillier, J. and Baker, R. F., 1944. *J. Appl. Phys.* **15**, 663.
Hillier, J. and Baker, R. F., 1945. *J. Appl. Phys.* **16**, 469.
Hillier, J., Baker, R. F., and Zworykin, V. K., 1942. *J. Appl. Phys.* **13**, 571.
Hillier, J., and Vance, A. W., 1941. *Proc. Inst. Radio Engrs., N.Y.*, **29**, 167.
Hofmann, U., 1943. *Wien. Chem. Z.* **46**, 97.
Husemann, E., 1943. *J. makromol. Chem.* **3**, 1.
Hutter, R. G. E., 1945 a. *J. Appl. Phys.* **16**, 678.
Hutter, R. G. E., 1945 b. *J. Appl. Phys.* **16**, 670.
Jakob, A. and Mahl, H., 1940. *Arch. exp. Zellfsch.* **24**, 87.
Kausche, G. A., 1940 a. *Arch. ges. Virusfsch*, **1**, 362.
Kausche, G. A., 1940 b. *Biol. Zbl.* **60**, 179.
Kausche, G. A., and Ruska, H. 1940. *Naturwissenschaften.* **28**, 303.
Keller, F. and Geisler, A. H., 1944 a. *J. Appl. Phys.* **15**, 696.
Keller, F. and Geisler, A. H., 1944 b. *Trans. Amer. Inst. Min. Egnrs., Tech. Publ.* No. 1700.
Klemperer, O., 1940. *Rep. Prog. Phys.* **7**, 107.
Knaysi, G. and Mudd, S., 1943. *J. Bact.* **45**, 349.
Kossel, W., 1943. *Naturwissenschaften*, **31**, 323.
Ladd, W. A., 1944. *Industr, Engng. Chem., Anal. Edn.*, **16**, 642.
Lembke, A. and Ruska, H., 1940. *Klin. Wschr.* **19**, 217.
Lembke, A., Ruska, H. and Christopherson, J., 1940. *Klin. Wschr.* **19**, 217.
Loring, H. S., Marton, L., Schwerdt, C. E., 1944. *Phys. Rev.* **65**, 354.
Luria, S. E., Delbruck, M. and Anderson, T. F., 1943. *J. Bact.* **46**, 57.
Mahl, H., 1939. *Z. tech. Phys.* **20**, 316.
Mahl, H., 1940 a. *Jahrb. A.E.G. Forsch.-Inst. Sonderheft* : *Übermikroskop*, **7**, 43.
Mahl, H., 1940 b. *Metallwirtschaft*, **19**, 488.
Mahl, H., 1940 c. *Z. tech. Phys.* **21**, 17.

MAHL, H., 1941. *Metallwirtschaft*, **20**, 983.
MAHL, H., 1942. *Z. tech. Phys.* **23**, 117.
MAHL, H. and PAWLEK, F., 1942. *Arch. Eisenhüttenw.* **16**, 219.
MAHL, H., and PENDZICH, A., 1943. *Z. tech. Phys.* **24**, 38.
MAHL, H., and STRANSKI, J., 1943. *Z. Metallk.* **35**, 147.
MARTIN, D., BAYLOR, M. R. B. and CLARK, G. L., 1944. *Science.* **99**, 185.
MARTON, C. and SASS, S., 1943. *J. Appl. Phys.* **14**, 522.
MARTON, C. and SASS, S., 1944. *J. Appl. Phys.* **15**, 575.
MARTON, C. and SASS, S., 1945. *J. Appl. Phys.* **16**, 373.
MARTON, L., 1934. *Bull. Acad. Belg. Cl. Sci.* **20**, 439.
MARTON, L., 1935. *Bull. Acad. Belg. Cl. Sci.* **21**, 553.
MARTON, L., 1936 a. *Bull. Acad. Belg. Cl. Sci.* **22**, 1336.
MARTON, L., 1936 b. *Physica.* **3**, 959.
MARTON, L., 1939 a. *Acta Union Internat. Contre Cancer*, **4**, 221.
MARTON, L., 1939 b. *Phys. Rev.* **56**, 290.
MARTON, L., 1940. *Phys. Rev.* **58**, 57.
MARTON, L., 1941. *J. Bact.* **41**, 397.
MARTON, L., 1943. *Ann. Rev. Biochem.* **12**, 587.
MARTON, L., 1944 a. *Phys. Rev.* **65**, 353.
MARTON, L., 1944 b. *Phys. Rev.* **66**, 159.
MARTON, L., 1944 c. *Science*, **100**, 318.
MARTON, L., 1945 a. *J. Appl. Phys.* **16**, 131.
MARTON, L., 1945 b. *J. Appl. Phys.* **16**, 387.
MARTON, L. and HUTTER, R. G. E., 1944 a. *Phys. Rev.* **65**, 161.
MARTON, L. and HUTTER, R. G. E., 1944 b. *Proc. Inst. Radio Eng., N.Y.*, **32**, 546.
MARTON, L., MCBAIN, J. W. and VOLD, R. D., 1941. *J. Amer. Chem. Soc.* **63**, 1990.
MARTON, L. and SCHIFF, L. I., 1941. *J. Appl. Phys.* **12**, 759.
MECKLENBURG, W., 1942. *Z. Phys.* **120**, 21.
MELNICK, J. L., 1944. *J. Immunol.* **48**, 25.
MRGUDICH, J. N. and CLOCK, R. C., 1944. *Trans. Electrochem. Soc.* **86**, 329.
MUDD, S., 1943. *Amer. J. Publ. Hlth.* **33**, 167.
MUDD, S., 1944. *J. Am. Med. Ass.* **126**, 632.
MUDD, S. and ANDERSON, T. F., 1944. *J. Amer. Med. Ass.* **126**, 561.
MUDD, S., HEINMETS, F. and ANDERSON, T. F., 1943. *J. Bact.* **46**, 205.
MÜLLER, H. O. and RUSKA, E., 1941. *Kolloidzschr.* **95**, 21.
NISSEN, H. F., 1944. *Z. Phys.* **122**, 573.
O'BRIEN, H. C. and MCKINLEY, G. M., 1943. *Science*, **98**, 455.
PICARD, R. G. and DUFFENDACK, O. S., 1943. *J. Appl. Phys.* **14**, 291.
PICHT, J., 1939. *Einführung in die Theorie der Elektronenoptik* (Leipzig : Barth).
PREBUS, A. F., 1942. *Engng. Exp. Sta. News, Columbus, O.* **14**, 6.
PREBUS, A. F., 1944. *Colloid Chemistry.* **5**, 152 (New York: Reinhold).
PREBUS, A. F. and HILLIER, J., 1939. *Canad. J. Res.* **17**, 49.
RAMSAUER, C., 1940. *Das freie Elektron in Physik und Technik* (Berlin : Springer).
RAMSAUER, C., 1941. *Elektronenmikroskopie* (Berlin : Springer).
REBSCH, R., 1938. *Ann. Phys., Lpz.*, **31**, 551.
RECKNAGEL, A., 1940. *Jahrb. A.E.G. Forsch.-Inst. Sonderheft: Übermikroskop*, **7**, 15.
RECKNAGEL, A., 1942. *Z. Phys.* **120**, 331.
RICHARDS, A. G. and ANDERSON, T. F., 1942. *J. N.Y. Ent. Soc.* **50**, 147.
RUSKA, E., 1934. *Z. Phys.* **87**, 580.
RUSKA, E., 1940. *Wiss. Veröff. Siemens-Konz. Werkstoffsonderheft*, 372.
RUSKA, E., 1942. *Arch. Elektrotech.* **36**, 431.
RUSKA, E., 1943. *Kolloidzschr.* **105**, 43.
RUSKA, E., 1944. *Kolloidzschr.* **107**, 2.
RUSKA, E. and MÜLLER, H. O., 1940. *Z. Phys.* **116**, 366.
RUSKA, H., 1939. *Naturwissenschaften*, **27**, 287.
RUSKA, H., v. BORRIES, B. and RUSKA, E., 1939. *Arch. ges. Virusfsch.* **1**, 155.
RUTHEMANN, G., 1941. *Naturwissenschaften.* **29**, 648.
RUTHEMANN, G., 1942. *Naturwissenschaften*, **30**, 145.

SCHAEFER, V. J., 1943. *Science*. **97**, 188.
SCHAEFER, V. J. and HARKER, D., 1942. *J. Appl. Phys.* **13**, 427.
SCHIFF, L. I., 1942. *Phys. Rev.* **61**, 721.
SCHMIDT, R. W., 1943. *Kolloidzschr.* **102**, 15.
SCHOON, TH. and BEGER, E., 1941. *Z. phys. Chem.* **189**, 171.
SCHOON, TH. and KOCH, H. W., 1941. *Kautschuk*, **17**, 1.
SCHULTZ, E. W., THOMASSEN, P. R. and MARTON, L., 1945. *J. Appl. Phys.* **16**, 265.
SEMMLER-ALTER, E., 1942. *Arch. Eisenhüttenw.* **16**, 223.
SHAW, B. T., 1942. *J. Phys. Chem.* **46**, 1032.
SIEGBAHN, K., 1943. *Ark. Mat. Astr. Fys.* **30**, 1.
SLIEPCEVICH, C. M., GILDART, L. and KATZ, D. L., 1943. *Industr. Engng. Chem., Industr. Edn.*, **35**, 1178.
STANLEY, W. M. and ANDERSON, T. F., 1941. *J. biol. Chem.* **139**, 325.
STARKS, H. J. H., 1935. *Rep. Prog. Phys.* **2**, 283.
SVARTHOLM, N., 1942. *Ark. Mat. Astr. Fys.* **28**, 1.
THOMASSEN, L., WILLIAMS, R. C. and WYCKOFF, R. W. G., 1945. *Rev. Sci. Instrum.* **16**, 155.
TISELIUS, A. and GARD, S., 1942. *Naturwissenschaften*, **30**, 728.
TURKEVICH, J., 1945. *J. Chem. Phys.* **13**, 235.
VANCE, W. A., 1941. *R.C.A. Rev.* **5**, 293.
WIEGAND, W. B. and LADD, W. A., 1942. *Rubber Age*, **50**, 431.
WILLIAMS, R. C. and WYCKOFF, R. W. G., 1944. *J. Appl. Phys.* **15**, 712.
WILLIAMS, R. C. and WYCKOFF, R. W. G., 1945 a. *Nature, Lond.*, **156**, 68.
WILLIAMS, R. C. and WYCKOFF, R. W. G., 1945 b. *Proc. Soc. exp. Biol. Med., N.Y.*, **58**, 265.
WILLIAMS, R. C. and WYCKOFF, R. W. G., 1945 c. *Science*, **101**, 594.
ZWORYKIN, V. K., 1943. *Trans. Amer. Inst. Min. Metall. Engrs.* **152**, 13.
ZWORYKIN, V. K. and HILLIER, J., 1943. *J. Appl. Phys.* **14**, 658.
ZWORYKIN, V. K., HILLIER, J. and SNYDER, R. L., 1942. *Bull. Amer. Soc. Test. Mat.* No. **117**, 15.
ZWORYKIN, V. K., HILLIER, J. and VANCE, A. W., 1941 a. *Trans. Amer. Inst. Elect. Engrs.* **60**, 157.
ZWORYKIN, V. K., HILLIER, J. and VANCE, A. W., 1941 b. *J. Appl. Phys.* **12**, 738.
ZWORYKIN, V. K. and RAMBERG, E. G., 1941. *J. Appl. Phys.* **12**, 692.

THE THEORY OF THE ELEMENTARY PARTICLES

By H. J. BHABHA, F.R.S.,
Tata Institute of Fundamental Research, Bombay

§1. HISTORICAL INTRODUCTION

THE concept of the elementary physical particles as a few types of simple physical entities which, by their interplay, generate the entire physical world may be said to have started its development in 1897, when the independent existence of the negative electron was established, notably by J. J. Thomson. The next important advance in the development was made in 1911, when Rutherford revealed the essential structure of the atom by his experiment on the scattering of α particles. By the middle of the third decade of this century, the picture of the atom as we know it now, though not including a knowledge of internal structure of the nucleus, had taken shape. It was known that the atom had a heavy nucleus whose radius was of the order 10^{-12} cm., and in which was concentrated all but about one part in two thousand of the total mass of the atom. The charge on the nucleus was invariably positive, and an exact integral multiple of the charge of the electron, while the masses of proximate nuclei differed by very nearly an integral multiple of the mass of the proton, the lightest nucleus known.*

This period marks a certain fairly definite stage in the development of the idea of the elementary particles. All known physical entities belonged to one of two types. On the one hand there were the electromagnetic and gravitational fields, whose state could only be described by giving the value of the field quantities at all points of space. On the other, there were the material particles, of which only two were elementary, the electron and the proton. All atoms and their nuclei were considered to be built up by putting together varying numbers of these two types of elementary particles. Moreover, it was believed that two particles of the same type were indistinguishable, and each one permanent and immutable. For example, the helium atom had two orbital electrons, while its nucleus, being about four times as heavy as the proton, was considered to be made up of four protons, the electric charge of two of these being neutralized by assuming two further electrons to be bound up inside the nucleus. The fact that the nuclei of certain radioactive elements emitted electrons was considered as evidence in favour of this picture.

Howoever, this simple picture was soon seen to require modification. First, the discovery of the fact that in certain circumstances material particles, like electrons, show some of the typical behaviour of waves, necessitated the well-known step from Newtonian mechanics to quantum mechanics. Conversely, the fact that in certain circumstances light shows some of the characteristics of discrete

* The mass of every nucleus is very nearly an integral multiple of one-sixteenth the mass of the oxygen (16) atom, and this integer is known as the mass number. As is well known, the mass of the proton is 1·00813 in terms of this unit, and hence the masses of heavy nuclei expressed in terms of the proton mass may differ considerably from the mass number.

particles necessitated the quantization of the Maxwell equations, a step which had been foreshadowed as far back as 1900 by Planck. Thus the sharp distinction between particles and fields had already become blurred, though, as we shall see presently, the complete fusion of the wave and particle aspects only takes place within the framework of relativistic field theory coupled with second quantization.

Similarly, the rapid development of theory, as well as the accumulation of further experimental knowledge of the nucleus, soon showed that all matter could not be thought of as built up from electrons and protons only. First, the formulation of the relativistic wave equation of the electron by Dirac in 1927 showed that it was theoretically impossible to bind an electron in a space as small as that of the nucleus, however strong the force might be. Secondly, the measurement of the spin of several nuclei and the type of statistics they obey showed that if there were electrons in the nucleus, these did not contribute either to the spin or the statistics. For both the electron and the proton have a spin angular momentum $\frac{1}{2}$ in units of Planck's constant $\hbar$ and obey the Fermi-Dirac statistics. Hence the deuteron, for example, which has a unit charge and twice the mass of the proton, should, if it were made up of two protons and one electron, obey Fermi statistics and have half integral spin, the orbital angular momentum due to the motion of these particles in a nucleus being always an integer. Actually the deutron has a spin 1 and obeys Bose-Einstein statistics. The same holds for the nitrogen nucleus of mass 14. The general experimental result is that the spin of the nucleus is half odd integral or integral, depending on whether its mass number is an odd or even integer, and the statistics is then respectively Fermi-Dirac or Bose-Einstein. Thus the assumption that the nucleus is made up of protons and electrons does not lead to agreement with the results of experiment.

The difficulty was resolved by the discovery of the neutron by Chadwick in 1932, a particle which has nearly the same mass as the proton, but no charge. One could then unify all the facts about nuclei on the assumption that the neutron also had a spin $\frac{1}{2}$ and obeyed Fermi-Dirac statistics, by supposing all nuclei to be built up of protons and neutrons only. The neutron must be looked upon as another elementary particle, for the same arguments as apply to nuclei prevent its being considered as made up of a proton and an electron.

The very admission of the neutron as an elementary particle, however, necessitated the abandonment of one of the properties hitherto attributed to an elementary particle, namely its permanence. For the existence of radioactive nuclei emitting electrons leads to the conclusion that in the act of β-emission a neutron must convert itself into a proton and the emitted electron be born. Since the spin of each of these particles is $\frac{1}{2}$ and the orbital spin can only be an integer, one must assume with Pauli that yet another electrically uncharged elementary particle, known as a *neutrino*, probably having a spin $\frac{1}{2}$, must be emitted with the electron. Moreover, the neutrino also allows one to assume that energy is conserved in the process of β-emission. For the energy states of the parent and newly formed nuclei are sharply defined and the change in energy of the nucleus is therefore fixed, whereas the electrons in β-decay are known to come out with a continuous range of energies less than this fixed value. One

may suppose the difference in energy to be carried off by a neutrino, which, because of its weak interaction with matter, eludes observation. The β-decay thus appears as the reaction

$$N \rightarrow P + e^- + n, \qquad \ldots\ldots(1)$$

where N, P, e^- and n stand for the neutron, proton, electron and neutrino respectively.

In relativity theory the energy E and momentum p, being the components of a vector, are connected by the formula

$$E^2 = p^2 + M^2, \qquad \ldots\ldots(2)$$

where M is an arbitrary constant which can be interpreted as the mass of the particle. (The velocity of light c can be taken to be unity by a suitable choice of the unit of time. Similarly, Planck's constant $\hbar$ may also be taken to be unity.) Thus for every value of the momentum there must correspond two values of E, and this feature must also occur in any relativistic quantum theory. As is well known, the Dirac equation allows states of the electron in which its total energy in a vacuum is negative. This difficulty, which would rob the theory of any physical meaning, was solved by Dirac by the assumption that all states of negative energy were already occupied, each with one electron. The further assumption that electrons obey the Fermi-Dirac statistics, as is known from experiment, now prevents any electron in a positive energy state from falling into a negative energy state. The (infinite) density of electrons in negative energy states which uniformly pervades the whole of space is assumed to be physically unobservable. Any departure from this regularity would, however, be observable. Thus, it can be shown quite easily that a "hole" made by removing one of the (negatively charged) electrons from a state of negative energy would behave exactly like an electron of positive energy and positive instead of negative charge. That the "hole" would behave like a particle of the same rest mass merely follows from the fact that the connection between the energy and momentum of the hole is the same as that for the electron whose absence creates it. Thus the necessity of interpreting the Dirac equation in a physically intelligible way leads to two important conclusions. First, the electron must obey Fermi statistics, for without the operation of Pauli's exclusion principle electrons in positive energy states are not prevented from falling into states of negative energy even if these are occupied. Secondly, electrons of positive charge, or *positrons* as they are now called, must exist in nature and be capable of being created in pairs (each with a negative electron) if enough energy is available for the purpose. Thus the Dirac theory for the first time explained a fundamental fact about the elementary particles, namely that particles of spin $\frac{1}{2}$ must obey the Fermi-Dirac statistics. It also provided for the first time an actual mechanism by which energy could be transformed into matter. For example, just as a quantum of light is absorbed by an electron in an atom, so a quantum whose energy was greater than $2M$ could be absorbed by an electron in a negative energy state, which would then jump to a positive energy state and become physically observable, at the same time leaving a hole behind which would manifest itself as a positron. In this process, which may be represented by the equation

$$h\nu \rightarrow e^+ + e^-, \qquad \ldots\ldots(3)$$

where e^+ stands for the positive electron, the energy of the quantum is converted into the rest mass and kinetic energy of the two particles.

The predictions of the Dirac theory have been remarkably confirmed by the discovery of the positron by Anderson in 1932, and by the observation of the process of pair creation in cosmic-ray showers by Blackett and others. Not only has the theory been quantitatively confirmed by experiment at energies of a few million volts, but the formulae for radiation loss and pair creation can be shown to be valid to the highest energies. For these processes form the basis of the cascade theory put forward in 1937 by Bhabha and Heitler and Carlson and Oppenheimer, and its predictions are in good agreement with experiments on the creation of cosmic-ray showers and the absorption of the soft component in cases where the energy of the particles and quanta involved is many thousand times the rest mass of the electron.

The Dirac theory brings about a further development of our concept of the elementary particles. The positron must certainly be considered as an elementary particle in the sense that it is not made up of other elementary particles. And yet its existence is indissolubly connected with that of the electron. One cannot make a physically satisfactory theory of electrons which does not at the same time lead to the existence of positrons and *vice versa*. One must therefore look upon both electrons and positrons as manifestations of one elementary physical entity, the quantized electron-field,* which in the course of its interaction with other physical entities takes on states which manifest themselves by the existence of a certain number of electrons and positrons.

We now come to the latest stage in the development of the theory of the elementary particles which is intimately bound up with the theory of nuclear forces. It has been found from a study of nuclei that the forces which act between any two nuclear particles have the following general characteristics: (1) The force is a short-range force much stronger than the coulomb force when the two particles are within a distance of about 2×10^{-13} cm. of each other, and falls rapidly to zero for larger distances, so that, for example, the nuclear force between the nuclei of two neighbouring atoms in a molecule is entirely negligible; (2) the magnitude of the force between like particles (two protons or two neutrons) and unlike particles (a neutron and a proton) appears to be of the same magnitude in corresponding states; (3) the force between a proton and a neutron appears to be an "exchange force" involving the transformation of the neutron into a proton and *vice versa* whenever interaction takes place.

The first successful effort to explain some of the features of nuclear forces was made by Yukawa in 1935. Starting from the fact that the equation

$$\left\{\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2} + \chi^2\right\} U = 0 \qquad \ldots\ldots(4)$$

has a static singular solution of the form

$$U = \frac{e^{-\chi r}}{r}, \qquad \ldots\ldots(5)$$

where $r = (x^2 + y^2 + z^2)^{\frac{1}{2}}$, a circumstance which had already been made use of

* This expression refers to the wave-field describing the electron, and must be distinguished from the electromagnetic field produced by the electron.

when Einstein introduced the so-called λ-term in his gravitation theory,* Yukawa assumed that the *nucleons* (a convenient word which is used to denote either a proton or a neutron) generate a field which satisfies equation (4). The resulting interaction between two nucleons is then a short-range one as required, and its magnitude can be adjusted by a proper choice of the value of the constant connecting the nucleons with the U-field. Equation (4) is, however, just the relativistic wave equation of a particle whose energy and momentum are connected by the relation (2) with $\chi\hbar$ written in place of M. It therefore describes a particle of mass $M=\chi\hbar$. The observed range of nuclear forces requires χ to have the value $0{\cdot}5\times10^{13}$ cm.$^{-1}$, and M has then a value of about 170 times the mass of the electron. (In anticipation of the later development, we shall straightway describe a particle of roughly this mass as a *meson*.) Further, in order to take account of the exchange character of the interaction between a proton and neutron, one must assume that the meson carries a unit positive or negative electric charge. Denoting the meson by μ, we thus have two processes,

$$\left.\begin{aligned} &\mathrm{N} \rightleftharpoons \mathrm{P}+\mu^-, \\ &\mathrm{P} \rightleftharpoons \mathrm{N}+\mu^+. \end{aligned}\right\} \qquad \ldots\ldots(6)$$

Yukawa fitted the β-decay into this picture by assuming that the meson had an interaction with an electron and a neutrino exactly analogous to (6). We write it in the more suitable form,

$$\left.\begin{aligned} &\mu^+ \rightleftharpoons \mathrm{e}^+ + \mathrm{n}, \\ &\mu^- \rightleftharpoons \mathrm{e}^- + \mathrm{n}. \end{aligned}\right\} \qquad \ldots\ldots(7)$$

By choosing the constant connecting the meson-field with the electron-field to be very small it is possible to make the theory give the rate of β-decay correctly. The process (1) must now be considered to take place in two virtual stages, namely, first (6) and then (7).

No elementary particle of rest mass intermediate between those of the proton and electron was known in 1935. But as a result of the clarification that had been introduced by the cascade theory in cosmic-ray phenomena, it soon became evident that a particle of this type must be present in cosmic rays, and by 1938 its existence had been definitely established. However, there was yet no experimental information about the spin of the meson, and as far as theory was concerned it could be either 0 or 1.

For a discussion of the meson theories of nuclear forces we refer the reader to a *Report* by R. Peierls (1939) in this series. One may sum up the present position by saying that the best account of nuclear forces is given by the theory of Møller and Rosenfeld (1939, 40). In this theory positively and negatively charged mesons of ***spin 0 and 1*** play a part, as well as neutral mesons. It will appear below that it is precisely with these assumptions that the best account of cosmic-ray phenomena can also be given.

The discovery of the meson introduces yet a new feature into our concept of the elementary particles. Since the mass of the meson is greater than that of

* It is interesting to note that the introduction of this term gives the gravitation quantum a rest mass $\hbar\sqrt{\lambda}$, which is now directly connected with the size of the finite (but unbounded) universe, since the latter is determined by λ.

the electron and neutrino, it is possible for the process (7) to take place in the absence of other matter. The meson must therefore be unstable, and it can be computed, by using constants chosen to fit β-ray data, that its average time of decay is of the order 10^{-8} sec. The spontaneous decay of the meson makes the meson a natural clock and provides a means of testing a prediction of the special theory of relativity in an extreme case. If τ is the mean time of decay at rest, then a meson of energy E must have a mean decay time $\tau E/\mu$, where μ is the rest mass (Bhabha, 1938). Experiments on the decay time of mesons of different energies have been made which confirm the slowing down of clocks in rapid motion in a striking way (cf. Euler and Heisenberg, 1938 ; Blackett, 1938; Rossi and Hall, 1941). Rossi and Hall find, however, that the mean proper lifetime of cosmic-ray mesons at sea-level is about $2{\cdot}4 \times 10^{-6}$ sec., a discrepancy with the calculated lifetime which had already been noticed earlier. Møller, Rosenfeld and Rozental (1939) have pointed out that the disagreement can be made to disappear on the basis of Møller and Rosenfeld's theory of nuclear forces by assuming different decay times for mesons of spin 0 and 1. Hamilton, Heitler and Peng (1943) have found that a quantitatively consistent picture of cosmic-ray and nuclear phenomena can be given by assuming that the primary cosmic rays are protons which in the very topmost layer of the atmosphere lose their energy by the rapid creation of mesons of spin 0 and 1. They further assume, following Møller and Rosenfeld, that mesons of spin 1 have a life of 10^{-8} sec., while mesons of spin 0 have a life of 10^{-6}. The latter constitute most of the hard component in the atmosphere, and it is their lifetime which is measured in experiments. The former decay almost at the moment they are formed, and are responsible for creating the soft component.

All the elementary particles so far known in nature have a spin of 0, $\frac{1}{2}$ or 1 unit.* The question now arises whether elementary particles of higher spin can exist. The only theoretical approach to this problem is to investigate whether consistent equations for particles of higher spin can be set up by taking as our guide some general principles underlying theories of particles of spin 0, $\frac{1}{2}$ and 1, which we may be reasonably certain are correct. The latest work on this question is reviewed in the next two sections.

As is well known, there are certain difficulties which still encumber the present field theories. They are of two types. First, there are the infinities which have an analogue in classical theory, as, for example, the infinite electromagnetic field energy of a point electron. Secondly, there are infinities of a specifically quantum nature for which there is no classical analogue. Such difficulties are, for example, concerned with what is known as the dynamical or transverse self-energy of the electron, and the fluctuation of the infinite zero-point energy of a quantized field. Considerable progress has been made in recent years in overcoming these difficulties, and this will be reviewed in § 4. The problems have not, however, been completely solved yet.

* The spins of the proton and neutron have only been observed in non-relativistic circumstances. The scheme put forward in § 3 (ii) would allow the proton or neutron to have any half odd integral spin greater than $\frac{1}{2}$, while, nevertheless, manifesting the spin $\frac{1}{2}$ only in their normal condition.

§2. PARTICLE-FIELDS OF SPIN 0, ½ AND 1

One of the primary requirements of every theory of the elementary particles is that it shall be in keeping with the principles of the special theory of relativity. Any mathematical quantities which appear in the equations must therefore have certain well-defined transformation properties for changes in the frame of reference, and these must be combined in the fundamental equations in such a way that the latter remain unaltered by any transformation of the Lorentz group involved in going from one Cartesian frame to another moving relative to it with uniform velocity. The only quantities of this description are tensors and spinors,* and one should expect that only those would occur in the theory of an elementary particle which are ***irreducible***, that is, which cannot be split into two or more parts in a relativistically invariant way. Correspondingly, we should expect the equations describing an elementary particle-field to be irreducible, that is, incapable of being split into separate sets of equations in a relativistically invariant way. Indeed, the fact that the elementary particles interact with each other and in some processes one type may disappear, giving place to another variety, as, for example, in the reactions (6) or (7) mentioned above, raises the question as to what percisely constitutes an elementary particle-field. One can answer this question as follows : *An elementary particle-field is one described by an irreducible set of equations*. In this sense the proton and neutron, despite the fact that they can transform into each other, must be regarded as the manifestation of two different particle-fields, while the positron and electron are manifestations of the same irreducible particle-field. A spinor has two types of indices which transform according to different inequivalent representations of the Lorentz group, and it is usual to distinguish these by putting a dot over indices of one type and leaving those of the other type undotted. We denote a spinor index by a Greek letter which can take on the values 1 and 2. Tensor indices will be denoted by Latin letters and take on the values 0, 1, 2, 3. Spinor indices can be lowered and raised just like tensor indices, and a repeated Latin or Greek index automatically implies summation over the values 0 to 3 in the former case and 1 and 2 in the latter. A spinor $a^{\alpha\beta\ldots}_{\dot\lambda\dot\mu\ldots}$ having k undotted and l dotted indices is irreducible if it is symmetric in all the k undotted indices and symmetric in all the l dotted ones. It has only $(k+1)(l+1)$ independent components, and it is said to transform according to the irreducible representation $\mathfrak{D}(k, l)$ of the proper Lorentz group. The transformation of the Lorentz group which merely reverses the direction of the three space axes turns it into the spinor $a^{\dot\lambda\dot\mu\ldots}_{\dot\alpha\dot\beta\ldots}$ with l undotted and k dotted indices, which accordingly transforms according to the representation $\mathfrak{D}(l, k)$. Both must appear in any equation for an elementary particle in order that it should be invariant under all transformations of the full Lorentz group including reflections. We denote the co-ordinates by x^k and take the metric tensor g_{kl} in the form $g_{00}=-g_{11}=-g_{22}=-g_{33}=1$, $g_{kl}=0$, $k\neq l$. A transformation t of the Lorentz group is one whose coefficients $t^k{}_l$ by definition satisfy the equation

$$g_{kl}=g_{rs}t^r{}_k t^s{}_l. \qquad \ldots\ldots(8)$$

* Tensor and spinor calculus are essential for this work, and their use cannot be avoided altogether. However, the reader unfamiliar with them will probably be able to pick up some general ideas by reading the text.

An infinitesimal transformation is given by $t_l^{\ k} = \delta_l^{\ k} + \epsilon_l^{\ k}$, where the $\epsilon_l^{\ k}$ are arbitrary small quantities and $\delta_l^{\ k} = 0$, $k \neq l$, $\delta_0^{\ 0} = \delta_1^{\ 1} = \delta_2^{\ 2} = \delta_3^{\ 3} = 1$. It follows from (8) that $\epsilon_{kl} = -\epsilon_{lk}$. This infinitesimal transformation is represented in any representation, irreducible or otherwise, by a matrix

$$\mathbf{T} = \mathbf{1} + \tfrac{1}{2} I^{rs} \epsilon_{rs}, \qquad \ldots\ldots(9)$$

where $\mathbf{1}$ stands for the unit matrix. The infinitesimal transformations I^{rs} completely determine the representation, and they satisfy certain commutation relations which only depend on the group and not on the representation concerned. These are

$$[I^{mn}, I^{rs}] \equiv I^{mn} I^{rs} - I^{rs} I^{mn} = -g^{mr} I^{ns} + g^{ms} I^{nr} + g^{nr} I^{ms} - g^{ns} I^{mr}. \qquad \ldots\ldots(10)$$

The simplest relativistic wave equation is (4), the so-called Klein-Gordon equation. It is obviously equivalent to the set of linear differential equations

$$\left.\begin{aligned} U_k &= \frac{\partial}{\partial x^k} U, \\ \frac{\partial}{\partial x_k} U_k + \chi^2 U &= 0, \end{aligned}\right\} \qquad \ldots\ldots(11)$$

as can be seen by eliminating the U_k from the second equation by using the first set. This equation describes a particle of spin zero since everything about the particle is known if the function U depending only on the co-ordinates is known.

The generalization of the Maxwell equations given by Proca (1936) is

$$\left.\begin{aligned} \chi G_{kl} &= \left(\frac{\partial}{\partial x^k} U_l - \frac{\partial}{\partial x^l} U_k \right), \\ \frac{\partial}{\partial x_k} G_{kl} + \chi U_l &= 0. \end{aligned}\right\} \qquad \ldots\ldots(12)$$

These have since been used to describe the meson. There are four wave functions U_k with one relation between them, namely $\frac{\partial}{\partial x_k} U_k = 0$, which can be deduced from (12), so that for every plane wave there are three different solutions. It therefore corresponds to a particle of spin 1 (in units of $\hbar$).

Finally, the electron is described by the Dirac equation, which we write in the form

$$(\alpha^k p_k + \chi)\psi = 0, \qquad \ldots\ldots(13)$$

where $p_k = -i \frac{\partial}{\partial x^k}$ and the α^k are four matrices satisfying

$$\alpha^k \alpha^l + \alpha^l \alpha^k = 2 g^{kl} \qquad \ldots\ldots(14)$$

As is well known, the relativistic invariance of (13) requires the α^k to satisfy the commutation relations

$$[\alpha^k, I^{lm}] = g^{kl} \alpha^m - g^{km} \alpha^l, \qquad \ldots\ldots(13\,a)$$

irrespective of whether the α^k satisfy (14) or not. The I^{rs} in this case are the infinitesimal transformations which determine the transformation property of the wave function ψ, and must satisfy (10). They can be interpreted as the operators connected with the spin of the particle. The algebra defined by (14)

has only one irreducible representation, namely, the usual one by matrices of four rows and columns.

Now both (11) and (12) can obviously be written in the form (13) if we take for the α^k matrices whose elements can be read off from (11) and (12) respectively. It was first noted by Duffin that in both cases the matrices so obtained satisfy the commutation rule

$$\alpha^k\alpha^l\alpha^m+\alpha^m\alpha^l\alpha^k=g^{kl}\alpha^m+g^{lm}\alpha^k. \qquad \ldots\ldots(15)$$

Conversely, Kemmer (1939, 43) has investigated the algebra generated by (15) and shown that there are only two non-zero irreducible representations, namely, those by square matrices of five and ten rows respectively. Thus (13) with the α^k satisfying (15) is mathematically equivalent to (11) or (12), depending on which representation is taken. This proof of equivalence, however, rests on taking a suitable representation of the matrices satisfying (15) and then showing that the resulting equations are (11) or (12). This procedure also has the disadvantage that the matrix method cannot be used in the calculation of processes involving interaction with nucleons, for there the field quantities such as U_k appear explicitly in the different interaction terms, and not ψ.

This lacuna has recently been filled by Harish-Chandra (1946 b). He has shown that when the α matrices satisfy (15), a matrix Γ_k of one row and four columns always exists such that $\Gamma_k\psi$ is an ordinary vector. It can be shown further that in the ten-rowed representation of the α's the Γ_k satisfy

$$\left.\begin{aligned}\Gamma_k\alpha_l&=-\Gamma_l\alpha_k,\\ \Gamma_k\alpha_l\alpha_m&=\Gamma_k g_{lm}-\Gamma_l g_{km}.\end{aligned}\right\} \qquad \ldots\ldots(16)$$

A transition from (13) to (12) is now immediately possible.

Defining two quantities U_k and G_{kl} by

$$\left.\begin{aligned}U_k&=\Gamma_k\psi,\\ G_{kl}&=i\Gamma_k\alpha_l\psi=-i\Gamma_l\alpha_k\psi=-G_{lk},\end{aligned}\right\} \qquad \ldots\ldots(17)$$

and multiplying (13) by Γ_l on the left, we get

$$-ip_kG_l^k+\chi U_l=0.$$

Multiplying (13) by $\Gamma_l\alpha_m$ gives

$$p_k\Gamma_l\alpha_m\alpha^k\psi+\chi\Gamma_l\alpha_m\psi=0$$

or, by (16),

$$p_mU_l-p_lU_m-i\chi G_{lm}=0.$$

These equations are just (12). An immediate correspondence between the "wave" formulation (12) and the "particle" formulation (13) can therefore be established, allowing a transition from one to the other at any stage of the calculation.

Precisely the same can be done for the five-rowed representation of the α's. The relations (16) have then to be replaced by certain others. The introduction of the Γ_k and a study of their algebra also allows general expressions to be written down immediately for the spurs ("traces") of the products of an arbitrary number of α's. The formalism should therefore prove of assistance in actual calculations.

It is important to note that in the case of all the spins considered above, the

infinitesimal transformations I^{kl} which determine the transformation properties of the wave function, and hence the spin of the particle, are given by

$$I^{kl}=[\alpha^k,\alpha^l]. \qquad \ldots\ldots(18)$$

Insertion of (18) into (13 *a*) shows that the α's satisfy

$$[\alpha^k,[\alpha^l,\alpha^m]]=g^{kl}\alpha^m-g^{km}\alpha^l, \qquad \ldots\ldots(19)$$

which is, therefore, a relation satisfied by the matrices for particles of spin 0, $\frac{1}{2}$ and 1.

It is fairly certain that the equations (11), (12) and (13) correctly describe the behaviour of particle-fields of spin 0, $\frac{1}{2}$ and 1. In order to develop equations for higher spin values one must find some general principles common to all of them. These are :—

A. It can be deduced from the equations that *each* component of the wave function satisfies the second-order wave equation (4). This is physically equivalent to the statement that the particle described by the field has in each case only one value of the rest mass (except for sign).

B. The particle-field is completely described by an equation of the form (13) without the help of any further subsidiary conditions.† The transformation properties of the wave function, and hence the spin of the particle, are determined entirely by the infinitesimal transformations I^{kl} defined by (18). Equivalently, the α matrices satisfy (19) for all spins.

§3. PARTICLE-FIELDS OF SPIN GREATER THAN 1

(i) *The Dirac-Fierz-Pauli scheme*

In 1936 Dirac gave a scheme of equations for particles of higher spin, taking as his guiding principle the postulate that the statement A shall hold for all spins. His equations connect a symmetric spinor $a^{\alpha\beta\ldots}_{\dot\lambda\dot\mu\ldots}$ in $2k$ undotted and $2l$ dotted indices, with a symmetric spinor $b^{\gamma\alpha\beta\ldots}_{\dot\mu\ldots}$ in $2k+1$ undotted and $2l-1$ dotted indices, where k, l are integers or half odd integers. The differential operator p_k in spinor notation takes the form $p^{\gamma\dot\lambda}$. Dirac's equations for a particle of spin $n=k+l$ are

$$\left.\begin{aligned} p^{\gamma\dot\lambda}\,a^{\alpha\beta\ldots}_{\dot\lambda\dot\mu\ldots}&=\chi b^{\gamma\alpha\beta\ldots}_{\dot\mu\ldots},\\ p_{\lambda\gamma}\,b^{\gamma\alpha\beta\ldots}_{\dot\mu\ldots}&=\chi a^{\alpha\beta\ldots}_{\dot\lambda\dot\mu\ldots},\end{aligned}\right\} \qquad \ldots\ldots(20)$$

and the equations obtained from these by a reflection. Either a or b can be eliminated from the equations, and it can be seen at once that each component separately satisfies the second-order wave equation (4). Further, contracting, say, the indices $\dot\lambda$ and $\dot\mu$ in the second equation, it immediately follows, on account of the symmetry of a in all its indices, that *

$$p^{\dot\mu}_{\gamma}\,b^{\gamma\alpha\beta\ldots}_{\dot\mu\ldots}=0. \qquad \ldots\ldots(21)$$

A similar equation holds for a. Now if we define another spinor $c^{\dot\delta\gamma\alpha\ldots}_{\dot\nu\ldots}$ by the equation

$$\chi c^{\dot\delta\gamma\alpha\ldots}_{\dot\nu\ldots}=p^{\dot\delta\dot\mu}\,b^{\gamma\alpha\beta\ldots}_{\dot\mu\dot\nu\ldots}, \qquad \ldots\ldots(22\,a)$$

* This follows from the fact that the spinors $\epsilon^{\dot\lambda\dot\mu}$ and $\epsilon_{\gamma\alpha}$ used for raising and lowering spinor indices are antisymmetric.

† See note added in proof, page 271.

then this spinor is symmetric in all its indices on account of (21), and we can immediately deduce from (22 a) that the equation

$$p_{\mu\dot{\delta}}\, c^{\dot{\delta}\gamma\alpha\ldots}_{\dot{\nu}\ldots} = \chi b^{\gamma\alpha\beta\ldots}_{\mu\nu\ldots} \qquad \ldots\ldots (22\,b)$$

also holds. (22) are therefore equivalent to (20). This process can be continued till one arrives at a spinor which has only dotted or undotted indices. There are, therefore, n different equations, if n is an integer, which are equivalent in the force-free case. For n half an odd integer there are $n+\frac{1}{2}$ sets of equivalent equations, which all describe a particle of the same spin n.

Fierz (1939) has quantized the scheme in the force-free case. It can be shown that for particles of integral spin the quantization can only be carried through so as to lead to Bose-Einstein statistics for the particles. On the other hand, for half-integral spins the total energy of the field is positive in states of positive mass and negative in states of negative mass. A physically sensible interpretation therefore requires that the field for particles of half odd integral spin shall be so quantized that the particles obey Fermi-Dirac statistics, in order that one may then overcome the difficulty about states of negative energy in the same way as in Dirac's electron theory. The essence of the connection between spin and statistics has been brought out by Pauli (1940). It has therefore been possible to find an explanation for the long suspected and empirically found connection between spin and statistics.

The Dirac scheme does not include all the equations which are consistent with the assumption A. For example, the scalar wave equation (11) for a particle of spin 0 cannot be written in the form (20). In fact, there is a corresponding set of equations for higher spins, namely,

$$\left.\begin{aligned} p^{\gamma}_{\dot{\lambda}}\, a^{\alpha\beta\ldots}_{\mu\nu\ldots} &= \chi b^{\gamma\alpha\beta\ldots}_{\dot{\lambda}\mu\nu\ldots}, \\ p^{\dot{\lambda}}_{\gamma}\, b^{\gamma\alpha\beta\ldots}_{\dot{\lambda}\mu\nu} &= \chi a^{\alpha\beta\ldots}_{\mu\nu\ldots}. \end{aligned}\right\} \qquad \ldots\ldots (23)$$

The Dirac scheme also has several serious disadvantages. The equations cannot be derived from a variation principle. This leads to the difficulty that when interaction with an electromagnetic field is introduced in the usual way by writing

$$\Pi_k = -i\frac{\partial}{\partial x^k} + eA_k \qquad \ldots\ldots (24)$$

in place of

$$p_k = -i\frac{\partial}{\partial x^k},$$

where A_k are the electromagnetic potentials, the equations become inconsistent, as has been shown by Fierz and Pauli (1939). The difficulty is in fact that the equations derived from a simple variation principle can only be reduced to (20) if the spinors satisfy the subsidiary conditions (21), and these do not come directly out of a variation principle. Moreover, these conditions are necessary in order that the statement A should be satisfied.

It has been shown by Fierz and Pauli (1939) that it is possible to derive the equations (20) as well as the subsidiary conditions (21) from a Lagrange function

by a special and rather complicated procedure. For example, a particle of spin $\frac{3}{2}$ can be described in Dirac's scheme by two spinors $a^{\alpha\beta}_{\mu}$ and $b^{\beta}_{\lambda\mu}$. To derive these equations from a variation principle Fierz and Pauli introduce two subsidiary spinors c^{α}, d^{α}_{μ} in the Lagrange function. The variation of all these spinors independently leads to 16 equations. The arbitrary constants in the Lagrangian are then so adjusted that in the force-free case these equations lead to the identical vanishing of the spinors c and d on the one hand and to the equations (20), with the subsidiary conditions (21), on the other. Interaction with an electromagnetic field can then be introduced by the usual procedure (24). In the presence of interaction, however, the subsidiary spinors do not vanish identically, but there are, nevertheless, just so many equations that for any given value of the momentum of the particle there are only four independent states, the number which a particle of spin $\frac{3}{2}$ must have.

The procedure becomes progressively more complicated for larger spins: for example, for a particle of integral spin n which can be described by a tensor $A_{k_1 k_2 \ldots k_n}$ symmetric in its n indices and satisfying $A^{l}_{lk_3 \ldots k_n}=0$, one has to introduce one subsidiary field $A_{k_1 \ldots k_{n-2}}$ of rank $n-2$, one of rank $n-3$, two of rank $n-4$, three of rank $n-5$ and so on to $n-2$ fields of rank 0. The Lagrangian is then built up as a sum of invariant terms quadratic in these fields multiplied by arbitrary coefficients. These coefficients are then adjusted so that as a result of the equations all the subsidiary fields vanish in the force-free case, and lead to the equations of Dirac together with the required subsidiary conditions for the principal field. Since the number of coefficients is larger than the number of conditions to be imposed, there remains an arbitrariness in the formulation which would manifest itself when interaction with other particle-fields is present.

It can also be proved (Bhabha, 1945 b) that for the equations of this scheme the statement B is certainly not satisfied. Moreover, this extended scheme has not been quantized, and it remains to be seen whether this can be carried through in a satisfactory manner.* In any case, we should expect a fundamental physical entity like an elementary particle-field to be described by equations which are simple and elegant, like those for particles of spin 0, $\frac{1}{2}$ or 1, and it therefore seems to me unlikely that the complicated scheme given above could be the correct one for describing particles of higher spin, if such exist.

(ii) *The alternative scheme*

An alternative scheme of equations was recently put forward by the author (Bhabha, 1945 a–45c) which takes as its point of departure the postulate that the statement B shall hold for all spins. The condition (18) plays a vital rôle now. One introduces a new index 4 and defines certain new quantities I^{k4} and g^{k4} by

$$I^{k4} \equiv -I^{4k} \equiv \alpha^{k}, \quad g^{44}=-1, \quad g^{k4}=g^{4k}=0 \qquad \ldots\ldots(25)$$

Then both (10), (13 *a*) and (18) can be combined into the one equation (10) by

* I have been informed by Prof. W. Pauli, in a letter, that the quantization of the scheme in the presence of electromagnetic interaction is involved and difficult, and that unpublished calculations by Weinberg and Kusaka exist on this subject.

letting the indices take on the values 0 to 4 instead of 0 to 3.* The resulting ten matrices I^{KL}, where a capital index runs from 0 to 4, are then seen to satisfy the same commutation rules as the nucleus of a representation of the Lorentz group in five dimensions, and the problem of finding an irreducible representation of the α's satisfying (19) can be solved immediately by using the theory of the representation of the Lorentz group in five dimensions, which is all known. *Each irreducible representation gives one irreducible equation for an elementary particle-field.* Every irreducible representation of the Lorentz group in five dimensions is characterized by two numbers n and m, which are both positive integers (or zero) or both half odd integers such that $n \geqslant m$. We denote this representation by $\mathfrak{R}(n, m)$. For this representation the eigenvalues of all the α's and I's (multiplied in some cases by i) are $n, n-1, \ldots -(n-1), -n$, which may occur more than once. Hence the four α's and six I's, multiplied in some cases by i, satisfy the algebraic equation

$$\{X^2-n^2\}\{X^2-(n-1)^2\}\{X^2-(n-2)^2\}\ldots =0, \qquad \ldots.(26)$$

which has just these roots. The last term is X or $X^2-\frac{1}{4}$, depending on whether n is an integer or half odd integer. This equation is of importance. Denote the operator which occurs in the wave equation (13) by $\mathbf{P}\equiv\alpha^k p_k$. The α's being matrices of finite degree, $\mathbf{P}$ must satisfy an algebraic equation of the same degree in which the p_k occur as coefficients. The Lorentz invariance of this algebraic equation now requires that the p_k should only occur in the invariant form $p^2=p^k p_k$. Now considering the equation in the system in which $p_0\neq 0$, $p_k=0$ for $k=1, 2, 3$, $\mathbf{P}$ reduces to $p_0\alpha^0$, and since α^0 satisfies the equation (26), $\mathbf{P}$ satisfies the equation

$$\{\mathbf{P}^2-n^2p^2\}\{\mathbf{P}^2-(n-1)^2p^2\}\ldots =0.$$

Operating on the wave function ψ with this equation, and replacing every $\mathbf{P}$ by $-\chi$, by a repeated use of the original wave equation (13), we find that every component of the wave equation satisfies

$$\{n^2p^2-\chi^2\}\{(n-1)^2p^2-\chi^2\}\ldots.\psi=0, \qquad \ldots\ldots(27)$$

the last term being χ or $\{\frac{1}{4}p^2-\chi^2\}$, depending on whether n is an integer or half an odd integer. The physical meaning of this circumstance is that the equation (27), and hence (13), has solutions in which energy and momentum are connected by the equation (3), with M given respectively, for integral or half odd integral n, by

$$\pm\frac{\chi}{n},\ \pm\frac{\chi}{n-1},\ \pm\frac{\chi}{n-2},\ldots\ldots\pm\chi \text{ or } \pm 2\chi. \qquad \ldots\ldots(28)$$

The particle-field therefore describes particles with several different values of the rest mass, but these must all be considered as different states of the same physical entity, just like the positron and electron, since the equations are irreducible.

For $n=\frac{3}{2}$ there are two different values of the rest mass $M=\frac{2\chi}{3}$ and $\frac{2}{3}M$.

* I have been informed by Prof. W. Pauli, in a letter, that this circumstance has been noted by Lubansky (1942). There appear to be other papers along somewhat similar lines by Kramers, Belinfante and Lubansky (1941) and Lubansky and Rosenfeld (1942). Since these were published in the Dutch *Physica* during the war, I have been unable to see any of them.

Only for $n = \frac{1}{2}$ or 1 is there a unique value of the mass, and the statement A is then also satisfied.

It can be shown (Bhabha, 1945 c) that the equations (13) can be derived from a Lagrangian

$$L = \psi^+ \mathbf{D}\{\alpha^k p_k + \chi\}\psi, \qquad \ldots\ldots(29)$$

where

$$\mathbf{D} = \text{const. } e^{i\pi\alpha^0}, \qquad \ldots\ldots(30)$$

these expressions holding for *all* spins. ψ^+ is the Hermitian conjugate of ψ. $\mathbf{D}$ commutes with α^0 and anti-commutes with the other three α^k. The algebraic equation (26) satisfied by α^0 depends on the maximum value of the spin n, and hence the Hermitian matrix $\mathbf{D}$ reduces to a polynomial in odd or even powers of α^0 with the leading term $(\alpha^0)^{2n}$. (30) reduces to the well-known expressions found by Dirac and Kemmer if $n = \frac{1}{2}$ or 1 respectively. Interaction with an electromagnetic field can now be introduced by the procedure (24). The current is given by

$$s^k = -e\psi^+ \mathbf{D}\alpha^k \psi \qquad \ldots\ldots(31)$$

and the energy tensor by

$$T^{kl} = \psi^+ \alpha^l p^k \psi. \qquad \ldots\ldots(32)$$

This tensor can be symmetrized as usual. The formal analogy with the Dirac equation of the electron is therefore complete.

The two numbers n and m, characterizing an irreducible representation $\mathscr{R}(n, m)$ of the Lorentz group in five dimensions, and hence a possible elementary particle-field, can be given a physical interpretation. n must be considered to be the maximum value of the spin of a particle described by the field, since it is the maximum eigenvalue of the spin operator. The physical significance of m becomes clear in the non-relativistic approximations of the equation. Thus, it can be shown that non-relativistically in the state of lowest rest mass the wave equation reduces to

$$\left\{\pi_0 - \frac{1}{2M}\pi_{\underline{k}}\pi^{\underline{k}} + \frac{e}{2nM} H_{\underline{l}} K^{\underline{l}}(m)\right\}\psi(n) = 0, \qquad \ldots\ldots(33)$$

where a bar below an index denotes summation over 1, 2 and 3 only. $M = \chi/n$ is the lowest value of the mass, H_l are the three components of the magnetic field and $K^l(m)$ are the three matrices representing the three components of angular momentum belonging to the total spin value m. The particle therefore behaves exactly like a particle of spin m and magnetic moment $e/2nM$ in its lowest mass state. m is therefore the spin of the particle in its stable state of lowest mass. The formula (33) includes at one stroke all the non-relativistic approximations that have so far been given for the electron and the meson, besides those for particles of higher spin in their state of lowest mass. For example, the Dirac equation is described in this scheme by $n = m = \frac{1}{2}$, and (33) shows that its magnetic moment is e/M. The Klein-Gordon equation is described by $n = 1$, $m = 0$, and (33) shows that its spin is zero. Finally, the vector meson equations are described by $n = m = 1$, and (33) then shows that the magnetic moment is $e/2M$ with a spin 1. Thus in this scheme of equations the different values of m for the same n describe physically different fields, unlike Dirac's theory for higher spins, where there are a number of mathematically equivalent alternatives for the same value of the spin n.

It has been pointed out (Bhabha, 1945 c) that the equation given by the scheme $\mathscr{R}(\frac{1}{2}, \frac{1}{2})$ describes a particle which behaves non-relativistically in its state of

lowest rest mass like a particle of spin $\frac{1}{2}$, as equation (33) shows, and it may therefore describe the behaviour of the proton. The experimental evidence so far at our disposal does not exclude this possibility, since all direct and indirect evidence of the spin of the proton has been obtained in non-relativistic conditions. Moreover, for this equation the particle has only one other value of the rest mass, and this is three times the normal value. Only in cosmic-ray phenomena is there enough energy to throw a proton into this state, and the chance of its observation even then would be very small because of its very short life. Other equations which are also not excluded for describing the proton are $\mathscr{R}(n, \frac{1}{2})$ (n half an odd integer). The number of different values of the rest mass is then $n+\frac{1}{2}$ and the ratio of the second lowest to the lowest value is $n/(n-1)$. Thus a very large value of n is certainly not permissible. The behaviour of the proton in the relativistic region would be different for all these equations, and it would be possible to decide in favour of one of these against the others by experiments on cosmic radiation, where it is believed that primary protons are responsible for creating mesons by a process analogous to the emission of radiation.

In the irreducible representation $\mathscr{R}(n, m)$ the α's have the eigenvalues n, $n-1, \ldots . -n$, which occur, say, $\nu_n, \nu_{n-1}, \ldots ., \nu_n$ times, the actual values of the ν's depending on the value of m. It is convenient to take α^0 in the diagonal form and bring equal eigenvalues next to each other by a rearrangement of the rows and columns thus:

$$\alpha^0 = \begin{bmatrix} n\mathbf{E}(\nu_n) & & & & \\ & (n-1)\mathbf{E}(\nu_{n-1}) & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & -n\mathbf{E}(\nu_n) \end{bmatrix} \quad \ldots\ldots(34)$$

where $\mathbf{E}(\nu)$ stands for the unit matrix of ν rows and columns. All the other elements are zero. $\mathbf{D}$ than takes the form

$$\mathbf{D} = -\begin{bmatrix} \mathbf{E}(\nu_n) & & & & \\ & -\mathbf{E}(\nu_{n-1}) & & & \\ & & \mathbf{E}(\nu_{n-2}) & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & (-1)^{2n+1}\mathbf{E}(\nu_n) \end{bmatrix} \quad \ldots\ldots(35)$$

For solutions in which the particle is at rest ($p_k = 0$, $k \neq 0$) ψ has the form

$$\psi = \phi e^{i\frac{\chi}{\hbar}t} \quad \ldots\ldots(36)$$

(where t has been written in place of x^0). The ϕ's are independent of the co-ordinates and describe the state of the spin. Substitution of (34) and (36) into (13) then shows that only those components $\phi(j)$ of ϕ are non-zero which correspond to the eigenvalue $-j$ of α^0 in (34). This solution describes a state in which the particle has rest mass χ/j and is at rest. Since the sign of $\mathbf{D}$ alternates in the successive squares in (35) we have the following situation. If n is an integer, the sign of $\mathbf{D}\alpha^0$ is opposite for the eigenvalues j and $-j$ of α^0, while if n is half an odd integer the sign is the same. Hence, for integral spins, for every state of rest mass χ/j there is another state with the opposite sign of j in (36) for which the charge density is opposite, while the total energy given by (32) is the same. The theory therefore leads to particles of positive and negative charge for each value of the rest mass exactly as in the case of spin 0 and 1. For half odd integral spins, for every state of rest mass χ/j there is another in which the charge is the same, but the energy opposite. Thus the same connection exists here between spin and statistics as exists for the equations of Dirac.

The quantum conditions for the wave function are

$$[(\psi^+(x')\mathbf{D}\alpha^0)_\mu, \psi^\nu(x)] = \delta^\nu_\mu\, \delta(x'^1 - x^1)\, \delta(x'^2 - x^2)\, \delta(x'^3 - x^3),$$

where, on the left-hand side, [·] stands for a commutation bracket for integral spins and an anticommutation bracket for half-integral spins. The quantization of this scheme requires a special and somewhat unusual procedure. Consider integral spins. In the state of lowest mass, χ/n, the energy of the field is always positive. In the next lowest mass state, $\chi/(n-1)$, however, the energy is always negative for both signs of the mass. This difficulty can be avoided as follows. When second quantization is introduced, one assumes that the metric in Hilbert space corresponding to the next lowest state of the mass is negative *definite* for all momentum states belonging to this mass value. Since a Lorentz transformation cannot change one mass value into another, this procedure is relativistically invariant. In the case of half-integral spins the difficulty is more serious. The alternation in sign of $\mathbf{D}\alpha^0$ has the effect that the coefficients of the Fourier components in the analysis of ψ belonging to the lowest but one value of the mass have to satisfy commutation relations of the type

$$-[b^* b + b b^*] = 1, \qquad \ldots\ldots (37)$$

which are mathematically impossible if b^* is the Hermitian conjugate of b. The mathematical difficulty can be overcome by adopting a procedure recently given by Dirac (1942, cf. Pauli, 1943) in another context. We assume b^* to be the adjoint of b with an indefinite metric in Hilbert space. With a proper choice of metric, b^* then becomes $-b^+$ or minus the Hermitian conjugate of b, and (37) can be satisfied. This procedure does *not* lead here to the same difficulties of physical interpretation as when it is applied to quantization in accordance with the Bose-Einstein statistics. For, due to the Fermi-Dirac statistics, in each state of a given momentum, energy and spin there can be either one particle or none, and one can so arrange the metric that the " negative probability " connected with an indefinite metric only occurs for the situation where no particle is present in that state. The negative probability then has no effect on the

charge or energy, since both are zero in this state. The theory therefore leads to physically reasonable results in every case.

The above discussion shows that the statements A and B are both fulfilled only in the case of spins 0, $\frac{1}{2}$ and 1. The postulate that they should both hold would therefore exclude any elementary particles of spin greater than 1. It has, however, been pointed out by Fierz and Pauli that the equations given by them for a particle of spin 2 and rest mass zero are exactly the same as Einstein's approximation to the gravitation equations for weak fields. Particles of spin 2 but zero rest mass do therefore appear to exist in nature (see also footnote on page 257 (§ 1) regarding the rest mass of the "graviton").

§4. THE DIVERGENCE DIFFICULTIES OF QUANTIZED PARTICLE-FIELDS

It now remains to consider briefly the limitations and difficulties of the quantized field theories for particles of spin 0, $\frac{1}{2}$ and 1. It is well known that the attempt to calculate a process to a higher approximation than the first one in which the process appears leads to infinities. In the case of the interaction of the electron with the electromagnetic field the second approximation is small compared with the first for most physical processes, and the difficulty is not serious. It remains, however, in principle and leads to such theoretically unsatisfactory results as an infinite self-energy of the electron. In the case of mesons and their interaction with nucleons, the higher approximations are not small compared with the first, and some solution of these difficulties is a practical necessity. As mentioned in the introduction, some of these difficulties exist also in classical theory, while others have no classical analogue. We consider them separately.

The classical difficulties are essentially of the following nature. The energy of the electromagnetic field produced by, say, an electric charge e spread over a sphere of radius r is of the order e^2/r, and, therefore, tends to infinity as $r \to 0$. The moment therefore that we deal with point electrons, the field energy must become infinite. It has, however, been pointed out (Bhabha, 1940) that this infinity has a physical meaning and corresponds to the fact that the work done in compressing a finite electric charge to a point is infinite. There is however no logical reason for regarding a point charge as the limit of a finite distribution, and it would be contrary to our concept of the elementary particles to think of them as made up in this way. It was first shown by Dirac (1938) that classically consistent equations of motion of a *point* charge can be found, and Pryce (1938) showed that the field energy could be redefined to agree with these equations. A general proof that this can be done for all types of fields and for point particles having a charge, dipole or higher multipole moment can be given, based on the inflow theorem (Bhabha and Harish-Chandra, 1945) which holds for every energy tensor which is conserved in empty space. Thus consistent equations for a point particle can always be found. For all fields satisfying the condition A or equation (4) it can be proved further (Bhabha and Harish-Chandra, 1946; Harish-Chandra, 1946 a) that a modified definition of the energy tensor can be given which is free from the worst singularities of the usual tensor. It can be shown that the equations for the particle so obtained can be put into Hamiltonian form by using the λ-limiting process of Wenzel and Dirac. Thus classically

the difficulties may be considered now to be completely removed with one exception—the methods do not immediately work in cases where the world lines of two particles cross.

Quantum mechanically, therefore, the λ-limiting process will remove all singularities where the corresponding classical singularities have been removed by it. It will not work in processes like pair creation (Pauli, 1943), but there is no difficulty in dealing with a process where the electric charge does not change, as, for example, in (6), where the charge of the proton is carried off by the newly-born meson.

Coming now to the specifically quantum-mechanical difficulties, these are due to the virtual emission and re-absorption of very energetic quanta, as, for example, in the transverse self-energy of the electron. An attempt to remove these has been made by Dirac (1942). His procedure requires the introduction of an indefinite metric in Hilbert space for certain momentum states of the radiation field. This has the result that those states in which an odd number of quanta are present in these momentum states appear with a " negative probability ", a circumstance which has no direct physical meaning. A different scheme for interpreting his equations has, therefore, been given by Dirac. But other difficulties remain in the Dirac scheme. Although the " negative energy " quanta cancel out the usual infinite transverse self-energy of the electron, they also cancel out several higher approximations which should be of physical significance. For example, Eliezer * has shown that higher-order corrections to the Klein-Nishina formula, which should represent the effect of the classical radiation damping, are also cancelled out.

Finally, attempts have been made (Heitler, 1941 ; Wilson, 1941) to derive from the usual perturbation equations of quantum mechanics another set which is devoid of the singular terms by simply leaving these terms out. The resulting equations then allow the calculation of higher-order processes. They thus allow radiation damping to be taken into account for processes involving the scattering of mesons by nucleons (Heitler and Peng, 1942). As far as scattering is concerned, the cross-sections obtained are essentially the same as those obtained on the basis of the classical theory with radiation damping (Bhabha, 1941). The quantum theory also permits the calculation of processes where one meson produces several others in collision with a proton.

To sum up, one may say that although it cannot yet be claimed that the problem of the divergence difficulties in the quantum theory of particle-fields has been completely solved, very considerable progress has been made in several directions and the theoretical position has definitely become clearer. One may reasonably hope that the remaining problems will be solved by a straightforward extension of the present formalism.

Our present picture of the physical world and the elementary particles may be formulated as follows. The permanent things in nature are certain generalized concepts like energy, momentum, angular momentum and electric charge, which are always conserved, while the actual elementary particles themselves are but the transitory embodiments of their metamorphoses. The underlying elementary particle-fields must, however, be conceded a permanent existence, for

* Unpublished work.

although at some given instant no particle of that field, as, for example, an electron, may be in existence, one could always demonstrate the continuing existence of the underlying particle-field by performing an experiment in which a particle of that type is created. In the equations for these particle-fields certain quantities enter as arbitrary mathematical constants, as, for example, the constants associated with the masses of the particles of the different fields, the electronic charge e, the " mesic charge " of a nucleon, and so on, which have quite definite values in nature. The explanation of why these constants have the empirically found ratios to each other appears to be completely outside the scope of the present scheme. To explain these ratios some radically new idea is required.

REFERENCES

Bhabha, H. J., 1938. *Nature, Lond.*, **141**, 117.
Bhabha, H. J , 1940. *Proc. Ind. Acad. Sci.*, A, **11**, 347–368.
Bhabha, H. J., 1941. *Proc. Roy. Soc.*, A, **178**, 314–350.
Bhabha, H. J., 1945 a. *Curr. Sci.*, **14**, 89–90.
Bhabha, H. J , 1945 b. *Rev. Mod. Phys.* (in course of publication).
Bhabha, H. J., 1945 c. *Proc. Ind. Acad. Sci.*, A, **21**, 241–264.
Bhabha, H. J. and Harish-Chandra, 1944. *Proc. Roy. Soc.*, A, **183**, 134–141.
Bhabha H. J. and Harish-Chandra, 1946. *Proc. Roy. Soc.*, A, **185**, 250–268.
Blackett, P. M. S., 1938. *Phys. Rev.*, **54**, 973–974.
Dirac, P. A. M., 1936. *Proc. Roy. Soc.*, A, **155**, 447–459.
Dirac, P. A. M., 1938. *Proc. Roy. Soc.*, A, **167**, 148–169.
Dirac, P. A. M., 1942. *Proc. Roy. Soc.*, A, **180**, 1.
Euler, H. and Heisenberg, W., 1938. *Ergebn. exakt. Naturw.*, **17**, 1–69.
Fierz, M., 1939. *Helv. Phys. Acta*, **12**, 3–37.
Fierz, M. and Pauli, W., 1939. *Proc. Roy. Soc.*, A, **173**, 211–232.
Hamilton, J., Heitler, W. and Peng, H. W., 1943. *Phys. Rev.*, **64**, 78–94.
Harish-Chandra, 1946 a. *Proc. Roy. Soc.* A, **185**, 269–287.
Harish-Chandra, 1946 b. *Proc. Roy. Soc.*, A (in course of publication).
Heitler, W., 1941. *Proc. Camb. Phil. Soc.*, **37**, 291–300.
Heitler, W. and Peng, H. W., 1942. *Proc. Camb. Phil. Soc.*, **38**, 296–312.
Kemmer, N., 1939. *Proc. Roy. Soc.*, A, **173**, 91–116.
Kemmer, N., 1943. *Proc. Camb. Phil. Soc.*, **39**, 189–196.
Madhava Rao, B. S., 1942. *Proc. Ind. Acad. Sci.*, A, **15**, 139–147.
Møller, C. and Rosenfeld, L., 1939. *Nature, Lond.*, **144**, 476–477.
Møller, C. and Rosenfeld, L., 1940. *Kgl. Danske Vid. Sels. Math.-Fys. Medd.* **17**.
Møller, C., Rosenfeld, L. and Rozental, S., 1939. *Nature, Lond.*, **144**, 629.
Pauli, W., 1940. *Phys. Rev.*, **58**, 716–722.
Pauli, W., 1943. *Rev. Mod. Phys.*, **15**, 175–207.
Peierls, R., 1939. *Rep. Prog. Phys.*, **6**, 78.
Proca, A., 1936. *J. Phys. Radium*, **7**, 347–353.
Pryce, M., 1938. *Proc. Roy. Soc.*, A, **168**, 389–401.
Rossi, B. and Hall, D. B., 1941. *Phys. Rev.*, **59**, 223–228.
Wilson, A. H., 1941. *Proc. Camb. Phil. Soc.*, **37**, 301–316.

Note added in proof (*see footnote to page* 262). It would be consistent with this standpoint of statement B to assume that the particle-field is completely described by an equation of the more general form

$$(\alpha^k p_k + \beta\chi)\,\psi = 0,$$

where β is an invariant matrix which commutes with all the I^{kl}. The connection (18) between the I^{kl} and the α's can then be generalized by making use of the matrix β.

THE MECHANICAL DESIGN OF PHYSICAL INSTRUMENTS

By A. F. C. POLLARD,
Imperial College, London, S.W.7

§1. INTRODUCTION

THE design of physical instruments may be divided into two main stages. The first is related to the theoretical planning of a particular instrument, and in so far as it is primarily concerned with the desired performance or functional operation of the instrument may be termed functional design. In the case of indicating or measuring instruments, the first consideration will be the specification of the permissible errors of reproducible indications before the theoretical development of the most efficient physical means for effecting the measurement is undertaken.

The second stage is concerned with the practical design, development and construction of the instrument mechanism, in conformity with mechanical principles and workshop processes.

Thus the design springing from a primary idea will assume, under the guidance of scientific principles and the methods and processes of practical construction, the form of the *first functional model*, in which the chief consideration is the desired performance of the instrument and the accuracy of the special service it is required to render.

During the development of the first functional model, every step in the process of design should be conducted with certainty, if waste of time and disappointment are to be avoided. To this end there must be no doubt that the several parts, when assembled, will conjointly function in the manner desired.

Only specially gifted persons are capable of directly carrying out the design of a complex mechanism without the incidence of mistakes, generally called " snags ". Apart from the necessity of being certain that a theoretical assembly of parts will function in the manner desired, the majority of minds are incapable of visualizing the interplay of the coacting parts incidental to their specified function, and other usually obscure effects upon the behaviour of the completed mechanism.

It is therefore advisable, indeed essential with complex mechanisms, to construct during the development of the design on the drawing board " trial " portions of the whole, the actual behaviour of which are beyond the imagery of the mind. Such precautions will shift the uneconomic redesign of a completed model to the simpler and more economic redesign of the lesser portions before final assembly of the whole.

There have been several attempts to elaborate systems of kinematic classification and notation to assist the imagery of the action and displacements of the component parts of mechanism when in action. The systems of Babbage (1826), Reuleaux (1876) and Hearson (1896) are particularly worthy of notice,

but from the time when Ampère (1834) first laid down the foundation of the science, which he named Kinematics, to the present day, no system has been devised of sufficient simplicity and width of scope to be generally applicable and of practical supplementary value to the kinematical theory of mechanisms. The involved complexity of Babbage's analytical engine, the drawings of which are stored in the Science Museum at South Kensington and run into hundreds of sheets, compelled him to devise his system, which in consequence is principally limited to gear mechanisms.

When assemblies of the component parts of a mechanism are a matter of kinematic investigation, the words used to describe them must have more precise meanings.

Such an assembly is conceived as consisting of movable *links* coupled together by pairs of *elements* as originally defined by Reuleaux (1876). Any link may consist of one or more components for purposes of construction, but in kinematics a link is a single rigid body. The element of a link consists of the functional surface or constraint of that link which mates with the element of an adjacent link.

The two elements thus coupled together form a *coupling*, and kinematic design is concerned with the design of the *couplings* which connect the links.

The assembly of coupled links will form a *kinematic chain* which may return upon itself and become *closed*, or which may be *open*, with two links at either end of the chain with uncoupled extremities.

A kinematic chain can serve no useful purpose unless one of its links is fixed in space. This condition must be fulfilled if the chain is to represent the kinematic elements of a *mechanism*. Kinematical analysis then becomes possible, and the real nature of the most disguised mechanisms becomes evident.

Actual mechanisms generally consist of many such chains compounded together, and the designer will be assisted in his conception of any mechanism if he draws a diagram of the whole assembly of compounded chains.

An important theorem at once flows from this conception of mechanisms, i.e. there are as many mechanisms as links in the chain or compounded chains.

The mechanism of an ordinary clock, for example, consists of many chains compounded together of which one link, consisting of the plates, is fixed. If now the hour hand be fixed, the rest of the mechanism revolves once in every twelve hours round its axis, and the mechanism becomes the recording drum of a barograph.

As far as functional design is concerned there is a real distinction between the first and second stages, but as far as material realization is concerned the distinction is ill-defined, since the one must be considered in view of the other in relation to workshop methods and processes which are common to both.

The common processes comprise the various methods of casting metals and other materials of construction—forging, turning, planing, shaping, milling and slotting, grinding, lapping, and honing and drilling, sheet metal pressing and spinning.

Each of these, to say nothing of the heat and other treatments of materials, has its practical limitations, which must be taken into account if impossible constructions are to be avoided.

For manufacture of a higher order, for interchangeable manufacture, other processes, such as drop forging, die casting, stamping, moulding of plastics,

involving the special design of the dies, and the gauging of limits of component dimensions, are pressed into service, and generally impose further limitiations upon design. In a short *Report* it is not possible to deal with all these processes and their development in recent years, or even the special precautions which must be taken when calling them into service, but a brief reference will be made ater to the universal process of casting and the precautions to be observed in the design of castings in so far as ordinary foundry practice and subsequent machining operations are concerned.

Apart from all other considerations, the design as a whole should be reduced to the greatest possible simplicity. Simplicity of the mechanism is not only of great economic importance in manufacture but also of undoubted functional advantage.

In addition, ease of assembly during manufacture is extremely important, for it is evident that unless the component parts of an instrument can be readily and quickly assembled, or dismantled, the economic gain in other directions may be jeopardized by fumbling at the assembly bench.

§2. THE SENSES IN OBSERVATIONS

When making observations with instruments, the sense organs are to be regarded as part of the instrument itself. The instrument does nothing more nor less than aid the senses. In other words, the errors due to the sense organs, as well as those due to the instrument, contribute to the errors incurred during observation, and for the purpose of design these two sources of error obviously must be distinguished and separated.

The sense organs not only suffer from functional imperfections, but are limited in their range of perception. The olfactory sense, for example, will detect minute quantities of certain substances, but it soon fatigues and, compared with some animals, and especially insects, is very limited in sensitivity and range of perception. Since the olfactory sense has received little attention, it may be worth while mentioning here that the results of some interesting researches on odours and the olfactory sense have been published by Moncrieff (1943, 44).

All the senses are subject to fatigue. More especially does the act of repeated accommodation fatigue the eye. In design, this disturbing factor is hardly ever remembered. Most measuring microscopes, for example, have scales or micrometer drums which must be focused by the eye in order to read them. But observation through the ocular requires the eye to be accommodated for infinity. In consequence, the spectroscopist, to take a case, must make repeated alternative observations of setting the graticule on the spectral lines and reading the micrometer, and he soon becomes thoroughly fatigued. If he persists, his personal errors of judgment in " coincidence " settings will very likely increase, and he will later make mistakes in reading the figures of the scale.

For similar reasons, it is important that instruments should be designed so that they may be used when the whole body of the observer is in a state of repose and in no way strained.

Most of our impressions are received through the visual sense, and the eye, with its neural apparatus, is the most highly developed sense organ we possess. Consequently, the great majority of instruments are visual instruments.

Without special knowledge of the limitations of vision and the manner in which the eye functions we are liable wrongly to interpret the impressions we receive through it, and it is advisable for the designer, as well as the research observer, to have access to Helmholtz's great work on Physiological Optics.

Though our most highly developed organ, the eye suffers from many imperfections which may affect our observations. The dioptric mechanism, for example, suffers from spherical and chromatic aberrations, and probably from other defects, such as astigmatism. The retina is limited in the range of its spectral sensitivity, both neurally and in virtue of the absorption of radiation by its dioptric elements. The resolution is poor and does not exceed one minute of arc or more with the normal eye, but under certain exceptional circumstances the resolution may be increased to 10 seconds of arc. It would be an idle procedure to try to improve, for instance, the mechanism of a coincidence range finder when the error of setting due to imperfections of the normal eye is greater than the error of setting or, rather, as will be seen later, the variance, due to the mechanism itself.

The field of observation at any instant is limited to the small projected area in space of the *fovea centralis*. Only one small area of an object can be seen clearly at a time, and much ingenuity, labour and money have been expended in attempts to produce aplanatic fields of view in the microscope, even though mechanical stages were previously developed to bring at will any part of an object quickly and precisely into the central small area of distinct vision.

The visual apparatus suffers, too, from persistence of vision, which for feeble illumination persists for about 1/10th second, but this interval can be reduced to 1/40th or 1/50th second with increase of illumination. We are quite unable to perceive the fluctuation of periodic impressions if the periodic time is less than the persistence interval.

It is evident that, armed with the senses alone, we should not be able to make very satisfactory progress in our attempts to observe and describe phenomena, as, indeed, the early history of science amply illustrates. Consequently we are compelled to devise instruments to aid or supplement the senses, and the importance of distinguishing the performance of our senses from the performance of those instruments which aid them cannot be too strongly emphasized.

A fundamental truth stated by Pollard (1938) is that "... the entire metrical fabric of the sciences is erected upon two ultimate experimental processes and on two only. We can count discrete, discontinuous quantities and we can compare two linear dimensions. As far as measurement is concerned we can do nothing more. Our counting can be accurately recorded by mechanism, but in the comparison of two lengths the instrumental aid to our senses terminates in a judgement. We judge by sight, by sound, by touch, or other sense, the approach of the two lengths to equality by a 'coincidence observation', and here we fail to make an exact measurement, for our senses are coarse and imperfect. Apart from the effects of errors whose sources are extrinsic to our instruments and ourselves, the exactness of the indications is not only limited by the range of the maximum variance, which is inherent in all instruments, but is also limited by the probability range of coincidence observations. We should endeavour therefore to reduce the former by correct design and the latter

by increase of sensitivity, until both lie within the limits of accuracy of the measured quantity we desire to attain.

"But there is a natural limit to the sensitivity of all instruments. The fluctuating zero of such instruments as highly sensitive radiometers and galvanometers had been observed years before Ising (1926, 31) pointed out the true nature of the indicator disturbances.

"So that as we increase the sensitivity of our instruments the particulate structure of matter and energy finally reduces the accuracy of our experimental comparison of two linear dimensions to the uncertainty of a statistical guess."

Ornstein and Milatz (1938), however, have shown that by causing the vibrating image of a line source from a fluctuating galvanometer mirror to darken a photographic plate, and by finding the point of maximum darkening by a photometric trace which gives the mean position of the line, then if the time of observation is increased by a factor p, the accuracy is increased by a factor $\sqrt{p}$.

It is important to remember in design that the errors due to coincidence observations may also be reduced, or rather "balanced", by applying identical methods of reading at the two extremities of the observed linear magnitude. This was first indicated by Rayleigh (1893) when describing the design of his manometer for measuring the pressure of gases. The measuring rod of the manometer terminates in two identical points directed downwards, and the length of the mercury column is compared with the distance between the two points by the usual method of surface reflection. By this means the errors of setting the two mercury surfaces is more likely to be the same in magnitude and sign than not and to balance one another. This will not be the case with the mercury barometer as usually designed, where the level surface of the mercury in the "cistern" is set by the reflection of a fixed ivory point, and the length between this surface and the top of the column is determined by the tangency of a knife edge with the curved meniscus of the mercury.

With the increasing mass and variety of measurements which must be made to check the numerous and varied processes in industry, and the necessity of reducing human error as much as possible, recording instruments have been developed. Though all records must be ultimately read by eye, recording instruments eliminate the necessity of constant observation, and in this sense may be said to supplement the senses.

When the response of the mechanism is neither indicated nor recorded, but used for purposes of control, we have a type of mechanism which, strictly speaking, is not an instrument, but stands in a class by itself. The moving element, equivalent to the indicator of an instrument, is used for the purpose of control; such, for example, is the needle of a carburettor or the sliding sleeve of a governor. Controlling mechanisms, however, are similar to instruments with respect to variance or non-reproducibility of the rest position of the controlling element.

Totalizing, aggregating or integrating instruments, which may be classed as counting mechanisms, usually operate in one sense, and their variance is of no importance.

But such instruments which indicate continuously and not in an intermittent

manner, such as planimeters, may operate in both senses during a single observation, and will suffer from variance, which should be taken into account.

Counting mechanisms register an accumulation of values up to the time of observation and may be kinetic or static.

Kinetic counting mechanisms, such as cyclometers, gas, water and electricity meters, and mechanical clocks, are discontinuous reading instruments in which the rest-position is positively defined by the physical intermittent contact of elements in the mechanism, and their variance is unimportant.

Other mechanisms which, strictly speaking, are counting mechanisms, having continuous motion due to the inertia of their moving parts, and always moving in one direction, are always in the cyclic state, and the question of variance does not arise. Electric clocks, for example, count the pulsations of the energizing current by operating in phase with it continuously, and are always moving in one direction.

Static counting mechanisms consist of links or components in the kinematic chain which are definitely located with reference to one another, and in design the coupling of these components may require to be carefully considered. Such instruments may give many or a single count. Thus in haemacytometers the cover-glass must always go back into the same position, so that the enclosed volume of diluted blood is accurate to within a specified error, to give the many counts of corpuscles accurately. A satisfactory way of ruling the reticule which does not influence the settling of the blood corpuscles has not yet been devised.

In specific-gravity bottles the stopper must go back into the same position with a similar specified condition, in order that, in the single count of a given volume, the volume may be constant within limits.

Volumetric glass-ware, in which volume is measured by observation of the coincidence of the meniscus with an engraved line, are measuring rods in which the compared linear magnitudes are functions of volume. Such instruments may suffer from variance to a significant degree and should always be checked from this point of view. No instrument can consist of a single link or component since, to function, that component must be supported by, or coupled to, a fixed link in order that it may function.

From what has been said, an instrument may be defined to be a mechanism consisting of links coupled together and designed to aid or supplement the senses. This definition divides instruments into two main groups, i.e. those which aid or supplement the senses in a qualitative manner and are used for observation only, such as microscopes, telephones, spectacles, and those which aid or supplement the senses in a quantitative manner, such as plug gauges, micrometers, voltmeters, range-finders.

The term scientific instrument is usually loosely used, but might with advantage be confined to those instruments which indicate or record a measured quantity.

The "scales" of indicating or measuring instruments may be linear or curved, and a "reading" consists of a comparison of the traverse of the indicator to the rest-position with the length of scale traversed. A more generalized meaning of the terms "scale" and "reading" will include any arrangement whereby the two lengths are compared. Thus it is logical to speak of the "scale" and

" reading " of a snap-gauge when used to compare the diameter of a cylinder with the fixed distance or " scale " between the measuring anvils by the sense of touch. With null readings, the actual procedure is to observe the approach of the indicator to zero reading, but principally on account of variance and, to a lesser degree, the error of " coincidence " judgment, the reading can never be effected exactly.

§3. MAXWELL'S AXIOMS

The primary requisites common to the construction of instruments, and indeed of all mechanisms, have been stated by Maxwell (1890), and are so important that his words should always be before the designer. He says: " There are certain primary requisites, however, which are common to all instruments, and which, therefore, are to be carefully considered in designing or selecting them. The fundamental principle is, that the construction of the instrument should be adapted to the use that is to be made of it, and in particular, that the parts intended to be fixed should not be liable to become displaced; that those which ought to be movable should not stick fast; that parts which have to be observed should not be covered up or kept in the dark; and that pieces intended to have a definite form should not be disfigured by warping, straining or wearing."

These apparently simple requisites of Maxwell may be regarded as the five fundamental axioms of design, one or other of which are frequently neglected, and sometimes are difficult to satisfy.

The first covers such defects as the want of necessary adjustments to secure the best performance of the instrument. Frequently the condensers of microscopes for example are not provided with centring adjustments. It also covers the inaccessibility of parts for assembly or adjustment. In some telescopes the objective is mounted in its cell so that, before " squaring on " can be effected, the objective, without which the adjustment cannot be made, must be removed. Again the inaccessibility of parts for cleaning is a common fault.

A violation of the second axion is illustrated by the loosening of nuts through vibration or other accidental causes, and the use of wrong-handed threads for nuts which secure pulleys upon their shafts. Almost unlimited ingenuity has been directed to the solution of the nut problem, as the work published by Schoeneich (1933) testifies, but no wholly satisfactory lock-nut has yet been devised.

" Parts which ought to be movable should not stick fast." The draw tubes of telescopes and microscopes sometimes stick and do not draw out smoothly. This is a fault of design, which should provide against the imperfections of tubes manufactured by draw-plate and triblet. These imperfections can readily be overcome by the application of the correct principles of design, which will be described later.

But an outstanding example of the violation of this axiom is that the cross-slides of the saddles of lathes sometimes jam on account of elastic deformation when the tool is clamped in position. This can easily be overcome by kinematic design.

Examples of a violation of the fourth axiom may frequently be found in optical instruments, and in some machine tools, in which a projecting part in certain

positions obstructs normal vision of a scale. The nose-piece of some microscopes, apparently added as an afterthought, obstruct the normal vision of the scales of the mechanical stage, and in some spectrometers the reading microscope cannot be used in particular positions of the telescope on account of the collimator.

It is practically impossible completely to satisfy the fifth and last axiom. The parts of all mechanisms are subject to elastic strain and wear. But elastic strain can be reduced to a minimum by correct design, and warping and wearing can be reduced by the proper treatment of the materials and surfaces used in the construction.

Maxwell (1890) continues:

"Each solid piece of an instrument is intended to be either fixed or movable, and to have a certain definite shape. It is acted on by its own weight, and other forces, but it ought not to be subjected to unnecessary stresses, for these not only diminish its strength, but (what for scientific purposes may be much more injurious) they alter its figure, and may, by their unexpected changes during the course of an experiment, produce disturbance or confusion in the observations we have to make.

"We have, therefore, to consider the methods of relieving the pieces of an instrument from unnecessary strain, of securing for the fixed parts a determinate position and of ensuring that the movable parts shall move freely, yet without shake.

"This we may do by attending to the well-known fact in Kinematics:

'A rigid body has six degrees of freedom'.

A rigid body is one whose form does not vary. The pieces of our instruments are solid, but not rigid. They are liable to change of form under stress, but such change of form is not desirable, except in certain special parts, such as springs. Hence, if a solid piece is constrained in more than six ways it will be subject to internal stress, and will become strained or distorted, and this in a manner which, without the most exact micrometrical measurements, it would be impossible to specify.

"In instruments which are exposed to rough usage it may sometimes be advisable to secure a piece from becoming loose, even at the risk of straining and jamming it; but in apparatus for accurate work it is essential that the bearings of every piece should be properly defined both in number and position."

In instruments designed in the ordinary engineering manner, the components nearly always suffer from some undesirable elastic strain, and in some cases it is quite useless to attempt to apply this type of "heavy design". A notoriously difficult problem is the mounting of very thin lenses in their cells, so that the elastic deformations do not warp the glass and alter the optical characteristics. But correct design, kinematic design, to which Maxwell alludes, can overcome these difficulties.

Spectacle lenses are sometimes so heavily strained by the encircling frame, or by the pinned-clamp method of fixation, that a jar will cause fracture. It is American practice to fit theodolites with four levelling screws instead of three. But when designed in this way the main socket of the centres can easily be overstrained, and numbers of such instruments are damaged in this way every year.

The castings and other parts of sensitive instruments constructed from pressed or rolled materials, such as interferometers and the capsules of aneroid pressure instruments, require careful treatment to eliminate internal strains, and the erratic behaviour of some instruments may be traced to unsuspected secular changes in the structure.

The annealing of most materials can never be completely effected, and the designer of sensitive instruments should be very suspicious of the behaviour with time of materials used in their construction.

The difficulty of completely annealing glass is well known, and much work has been done in attempting to improve annealing processes for this material.

With standards of length the greatest care must be exercised in the selection and treatment of the materials from which they are constructed.

Rolt (1929) states that the alloy invar is unstable in the annealed as well as the unannealed state, the instability taking the form of a gradual increase in the length of a bar with time.

An annealed invar metre-bar, which has been under observation at the National Physical Laboratory for about 25 years, increased its length by 0·024 mm. during that period. The rate of increase in length is diminishing, but the bar was still growing at the end of this period at the rate of about 0·0004 mm. per annum.

A new type of invar with a smaller secular rate of growth has been developed by Guillaume (1927).

A third nickel-iron alloy, with a coefficient of linear expansion agreeing closely with the average figure for steel, is reported to be quite stable.

Probably crystalline or even fused quartz is the most stable material with which we are acquainted. Rolt (1929) states that a metre standard of fused quartz at the National Physical Laboratory has been shown to be remarkably stable over a period of 17 years.

Pure nickel is also highly stable. The nickel standard metre at the National Physical Laboratory has been found to remain constant in length to within $\pm 0{\cdot}1$ micron over a period of 16 years. But this metal has a high coefficient of linear expansion.

The unexpected manner in which cast or wrought works shrink and change their shape should be of interest to the designer of sensitive instruments, and he should be acquainted with the classical experiments of Lt. Col. H. Clerk (1863) and Thomas Wrightson (1879), the remarkable results of which aroused the interest of Sir George G. Stokes.

The extensive shrinkage and peculiar behaviour of cast iron on cooling requires castings to be designed so that every part shall cool at about the same rate, if excessive strains and probable fracture are to be avoided. A noteworthy case of the persistent bursting on cooling of a wrongly designed casting has been recorded by Pollard (1933 a).

Only brief reference can be made to these matters here.

The kinematical theorem, i.e. a rigid body has six degrees of freedom, to which Maxwell refers, when properly applied to the design of the coupling between the component parts of mechanisms not only reduces the unnecessary strains to a minimum and correctly defines the bearings of every piece, both in number

and position, but greatly increases the reliability of repeated readings and eliminates in manufacture the costly and troublesome limits of fit of mating surfaces, which are necessary when couplings are designed on the usual engineering principles. This type of design formerly called geometric has been more correctly termed kinematic by Pollard on account of the precision of the relative motion of the two coupled components it is possible to control by its application.

§4. THE ERRORS OF MEASUREMENTS

Since the principal object of the functional design of an instrument is to reduce the errors of indication to a minimum, or to limit them to specified amounts, it is advisable to distinguish the various kinds of error which may arise in measurements.

Jeffreys (1931) has clearly stated the significance of measurement and the nature of our deductions from it. When making measurements it is assumed that the true values exist, but we must avoid regarding the observed values as anything but approximations to the true values.

In practice the true values, not being known, do not exist, and a compromise has to be made. If we have a number of observations of one or more variables, a simple law is found to fit them approximately. The law may involve some parameters not already known, and it will usually be impossible to make the law fit all the observations exactly.

Parameters may be chosen, however, to fit the observations as closely as possible, though it is a matter of convention what criterion is adopted to measure the closeness of fit.

When such a criterion has been chosen, the parameters in the law and the values of the function are unique. These are called the adopted values. The adopted values, however, may, and in general will, differ from the assumed true values, but will be a closer approximation to the true values than the observed values.

The difference between the observed and adopted values are generally termed the "residual errors" or "residuals", and the difference between the true and adopted values may be called the errors of the adopted values.

The adopted values exist in our conceptual field and not in the external world.

It is assumed that the exact simple law, which the observations fit approximately, refers to certain specifiable conditions, and the residuals arise from the fact that the natural conditions of the experiment differ to some extent from these ideal conditions.

The practical justification for these assumptions is that we find the more closely these ideal conditions are realized the more accurately the simple law fits the observations, though it never fits them precisely.

These ideal conditions, which can never be attained, refer to conditions intrinsic as well as extrinsic to the instrument itself, a distinction which must be made with great care in design.

However carefully, then, a measurement is made, it is to be regarded as subject to error. This is especially the case when the limits of our measuring power are approached, and the repetition of the observations a number of times will not yield the same results.

Errors of measurement may be due to faults of construction or errors

of graduation of the instrument, to changes in the external conditions, i.e. in such parts of the environment of the observer as effect either him or his instrument, and are beyond the observer's control, or to the personal peculiarities of the observer. Such errors will be of a more or less systematic nature, and we might expect that when the observations have been corrected for all these sources of error, the corrected observations would fit the simple law exactly.

But in practice, when a series of observations is made and corrected for all known sources of error, it is found that the individual corrected measurements differ from one another. These differences are attributed to a class of error which is called an accidental error, due to variations in conditions of observation which have not been recognized as variable.

The evaluation of the most plausible value of any quantity involved in the measurement then requires, as previously stated, the adoption of a criterion for the discussion of the observations. The most widely used criterion is that the sum of the squares of the residuals is a minimum. In the simplest case, where a single quantity is measured, this criterion leads to the result that the adopted value is the mean of all the measurements.

Other criteria have been devised for the reduction of observations, and in some cases they give results differing appreciably from those obtained by the method of "least squares". But once a definite criterion has been adopted, the subsequent reduction of the observations is impersonal and leads to a unique solution of the problem of determining the best values or the reduction of the errors of the adopted values to a minimum.

Accidental errors usually conform, at least approximately, to the normal or Gaussian law of frequency distribution, and since this frequency distribution is symmetrical, the accidental errors may conveniently be called symmetrical errors.

Mellor (1916) gives references to cases, however, in which the number of large errors actually found is in excess of theory.

When the observation involves two or more parameters, the values of which are to be found, the method of least squares is applied to compute their values. Masterly and lucid accounts of the methods of computation to be used in such cases have been given by Chauvenet (1891) and by Brunt (1931).

That class of errors usually termed constant, regular or systematic are those produced by the persistent influence, for a short or long time, of disturbing factors. Such errors, more often than not, are not symmetrically disposed, and for purposes of distinction may be classed as asymmetrical errors. This class may then include irregular errors as well as truly systematic errors.

The causes of the systematic errors are generally known and are subject to correction, but the causes of the irregular errors, which tend to be unidirectional or unsymmetrical, are generally erratic and obscure and too complex to analyse.

Irregular errors, for example, are characteristic of mercury-column indicating instruments, in which the chemical condition of exposed surfaces of the mercury during rest and motion may change with the time and cause changes in the surface tension and frictional drag over the containing surfaces in a unidirectional but erratic manner. This phenomenon is probably due to the rupture and rapid retormation of surface films.

Quincke found that the angle of contact of a freshly formed drop of mercury on a glass plate steadily diminished with the time. Thus the angle of a fresh drop was 148° 55′, but, thereafter, steadily diminished. After two days it was only 137° 13′. On tapping the plate the angle rose to 141° 19′, and after another two days fell to 140°.

Gore (1879) found that the capillary electrometer behaved in an erratic manner, and other investigators (Turner, 1937) have found that the displacement of the mercury in sensitive thermometers is erratic. Interesting cinematograph studies of the irregular movement of mercury in open-scale thermometers were exhibited at the Physical Society's Annual Exhibition of Scientific Instruments and Apparatus in 1933.

Erratic behaviour of other liquid-column instruments is known and the causes are obscure. The marked passiveness of a level vial has been reported by Starling (1922–23), for which no explanation could be given. The irregular errors due to such causes, as previously stated, are frequently biased in sign, and their effect is to produce an unsymmetrical frequency distribution, when the application of the ordinary least-square methods is then meaningless.

The possible presence of irregular errors in observations with instruments makes it essential to reduce the instrumental variance which may include such errors to such an extent as design and workmanship will permit.

These different errors of observation, and all those peculiar to instruments, do not appear to have been clearly distinguished in the literature. Consequently, it may be helpful to class them as follows:

1 Symmetrical errors, or so-called accidental errors.
2 Asymmetrical errors.
2.1 Systematic or constant errors.
2.11 Theoretical corrections.
2.12 Instrumental variance.*
2.13 Personal corrections or personal equation.
2.2 Irregular errors.

As indicated, the systematic errors may be separated into those which may be computed, those due to the *variance* or energy losses in instrument mechanisms, to be described later, and those due to peculiarities of the observer. Theoretical corrections of errors produced by known causes, such as temperature changes in the length of measuring rods, errors of graduation of scales and of eccentricity of divided circles, of the pitch of micrometer screws, present no difficulties of comprehension and of correction to at least the first order of small quantities.

For example, most of the errors due to the eccentricity of divided circles, as well as the greater part of the effect of the periodic errors of graduation of the scale, can be eliminated by reading an even number of verniers, say four or six, placed at equal distances around the circle (Ball, 1908).

It is of course desirable that the circles of precision instruments, such as astronomical instruments, should be engraved with the greatest possible accuracy.

* The word "variance" is used in statistics to denote (standard deviation)^2. Unfortunately this word is more descriptive of this class of error than the word variability, which has been suggested. Perhaps confusion can be avoided by the qualification "instrumental", as suggested above.

The accuracy of some circles is remarkable. Even previous to 1858 dividing machines must have reached great perfection, for Struve is quoted by Sir John Herschel (1858) as having stated that the probable amount of the accidental error of division of the great Ertel circle at Pulkova does not exceed 0·264 second. This circle presumably is the one constructed by Ertel and Sons, Munich, with a diameter of 43 inches, divided to 2 minutes of arc and read by four microscopes, which read directly to single seconds and by estimation to 0·1 second or even less. The construction of the master circle of dividing machines is notoriously difficult and tedious, but by the method of determining the errors of circles by reference to an imaginery perfect circle, as described by Bouasse (1917), an almost perfect circle could be engraved by reference to an imperfect circle.

The personal errors, which as a rule are asymmetrical, are difficult to discover and may remain constant for months or even years. As stated by Brunt (1931), a careful and experienced observer appears to commit an error which is generally of the same sign, and approximately of the same magnitude, in a series of similar observations. An inexperienced observer is liable to commit errors of varying magnitude and sign, and even an experienced observer, when tired or not in his normal state, ceases to have a definite personal equation.

The literature on the personal equation has greatly increased since Arago (1853, 59) distinguished and described this class of error, which arises from imperfections of the senses, previously referred to, and from variable delays in response of the neural mechanism.

Instrumental variance, a defect from which all instruments must necessarily suffer, is the most vital of all classes of error to the designer of instruments. Except in unusual instruments, variance greatly exceeds in magnitude the symmetrical errors which are superimposed upon the two branches of the hysteresis loop defining the variance at every point of the indication range.

§5. FUNCTIONAL CHARACTERISTICS OF INDICATING INSTRUMENTS

There are five functional characteristics of all indicating instruments, which have been clearly defined and expressed quantitatively by Schlink (1918). These are: *accuracy*, *precision*, *sensitivity*, *passiveness* and *instrumental variance*.

The distinction between accuracy, sensitiveness and passiveness has not always been carefully maintained in the discussion and calibration of physical and engineering indicating instruments.

Variance is a characteristic of the greatest importance which had not been discussed previously to Schlink (1919 a, 19 b).

(i) *Accuracy*

It is generally assumed that an instrument is accurate when its indications accord with the values of the measured quantity. This is an ideal which, principally on account of the dissipation of energy in all mechanisms during motion, can never be realized. Consequently the readings never accord with the measured quantity, and in the process of calibration the errors of the readings are discovered and applied as corrections. These errors may arise from any cause whatever, and may be changed by changes in the environment and by

adjustments of the instrument mechanism. Thus by altering the tension of a restoring spring, by altering the length of some link or the contour of a cam, or by some other alteration in the mechanism such as the spacing of graduations, the magnitude of these errors may be changed. They have been classed above as theoretical corrections. The *accuracy* of an indication is inversely proportional to the error, and the *specific accuracy* is defined as the ratio of the measured quantity to the error of the indication.

The reciprocal of the specific accuracy may be termed the *specific inaccuracy*.

These specific quantities are not useful in the ordinary use of the instrument, but will be of service in rating the performance of an instrument.

Generally, industrial indicating instruments are not so sensitive as laboratory instruments, and are designed to dispense with the necessity of applying corrections. But with laboratory instruments it is generally supposed that such corrections are necessary if reliance is to be placed upon the indications. In the light of the inevitably present variance, however, the validity of this procedure is open to question.

(ii) *Precision*

Precision is associated with accuracy by many writers, but it is desirable to associate this term with the literary meaning of exactness or the negation of vagueness or looseness. An instrument may have more precise indications at one part of the range of its indications than at another. Precision refers to the extent of scatter of errors about the median, and is directly proportional to the probable error, the mean error and the error of mean square of a number of measures of an observation. It implies the quality of close limit to repetition indications and is contained within the variance. The smaller the variance at any value of the measured quantity the greater the precision of a random indication. The modulus of the normal frequency-distribution is the numerical measure of precision.

(iii) *Sensitivity and passivity*

The term sensitivity is frequently used in an unscientific sense, implying that a change in the quantity being measured is accompanied by a change in the indication of the instrument. For particular instruments, or those read by null methods, sensitivity is expressed in terms of the smallest change in the measured quantity which causes perceptible response in the indicator. Analytical balance makers, for example, more often than not, specify the sensitivity of their instruments as "turning" to so many milligrammes or fraction of a milligramme. Obviously the beam might be sluggish and might turn to a smaller difference of load than that stated if left to itself long enough or if exposed to imperceptible vibration.

Expressed in this way, sensitivity would also depend upon the observer's acuteness of perception and would result in estimates of the sensitivity, which depend upon circumstances extrinsic to the instrument, whereas sensitivity is an inherent and constant characteristic of the instrument itself.

The sensitivity, at any indication of the instrument, is the rate of displacement of the indicating element with respect to the measured quantity.

Thus in figure 1, if the curve gives the relation between the indication θ of the instrument at every point of the scale range and the value m of the measured quantity, the sensitivity, at any reading θ is $\frac{\delta\theta}{\delta m}$, where $\delta\theta$ and δm are the small intervals between two consecutive readings taken close together. In the limit the sensitivity becomes the slope of the curve at any value of the indication.

On account of variance, however, the sensitivity of all instruments must have two values, corresponding to the slopes of the two branches of the hysteresis loop. The single curve referred to above, when defining the sensitivity, is intended

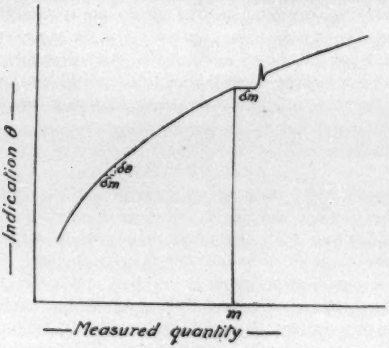


Figure 1.
(*Reproduced by courtesy of Adam Hilger, Ltd.*)

to be the curve lying midway between the two branches which will yield the ***mean sensitivity*** of the instrument at any given indication. In the preliminary stages of design, when specifying the sensitivity, it can only be the mean sensitivity which is implied, since the hysteresis loop is indeterminate until the instrument is constructed.

The utility of investigating this theoretical sensitivity may be illustrated in the case of the common spherometer mentioned on p. 295.

Here h is the indication and ρ the measured quantity; hence the sensitivity is

$$\frac{dh}{d\rho} = -2\left(\frac{h}{R}\right)^2,$$

neglecting h^3 compared with R^3.

If, now, h be kept constant, and the diameter of the contacting ring were made to vary and to serve as the scale of the instrument, the sensitivity would be

$$\frac{dR}{d\rho} = \frac{h}{R}.$$

The ratio of the latter to the former sensitivity is $-\frac{1}{2}\frac{R}{h}$, so that if $R=2$ cm. and $\rho=20$ cm., $h=0{\cdot}1$ cm. and the ratio $=10$.

This at once makes us suspicious of the soundness of the existing design where h is the indicator, as compared with a design in which R is the indicator.

It should be noted that this definition of sensitivity, as the slope of the indication/measured-quantity curve, does not imply that the sensitivity at any given reading can be determined by merely noting two readings and the corresponding values of the measured quantity taken close together. In instruments of less precise character a considerable change of the measured quantity may be effected without any initial response in the indicator. This period of non-response at the commencement of the motion of the mechanism is due principally to backlash, static friction or the elastic yielding of co-acting parts.

Previous to the work of Schlink (1918), this sluggishness was not clearly distinguished from insensitiveness, and he named it passiveness.

Passiveness, then, may be determined by observing the smallest alteration in the quantity to be measured which will produce any change in the indication of the instrument. The amount of the least alteration in the value of the measured quantity producing response of the indicator divided by the initial value of the measured quantity may be termed the *passivity* of the instrument at that point.

Though any association of "perceptible movement" with the notion of sensitivity has been rejected, it is permissible with passiveness, since the period of non-response has been overcome as soon as any motion whatever of the indicator takes place. In figure 1, the curve at the initial value m becomes horizontal over a range δm, and the end of this range is indicated by a small damped oscillation of the indicator, in order to bring clearly before the mind the period of inactivity of the indicator, which, being overcome by the growing force acting upon it as the measured quantity increases, is suddenly released and oscilllates about its new position of equilibrium, corresponding to the increased value of m. The passivity in this case is $\frac{\delta m}{m}$.

Passiveness is generally always present at the beginning and end of the complete hysteresis loop of the instrument, where a reversal of motion takes place, and is, therefore, contributory to the variance.

(iv) *Instrumental variance*

Let any indicating instrument, such as an ammeter, a micrometer microscope or theodolite, be taken carefully *in one direction* through the range of its indications, and the indications at intervals be observed against the value of the measured quantity, and the same operation be repeated in the opposite direction through the range of indications back to the initial value. Then it will be found that if the instrument indications be plotted against the values of the measured quantity, the two curves thus obtained will not coincide, but in general will form a closed loop, which is termed the *hysteresis loop* of the instrument.

If the ascending and descending branches of the hysteresis loop have been carefully determined by strictly aperiodic variation of the indications in this way, and the environment has not changed during the experiment, the boundaries of the loop will be the extreme variations of reading for any value of the measured quantity, and any subsequent reading afforded by the instrument, under the same

conditions of environment, will fall within these limits for any value of the measured quantity. This range of uncertainty of indication, in that it represents the range within which random readings of the instrument may be expected to lie, when all causes of error except those inherent in the instrument itself are eliminated, is conveniently termed the *instrumental variance*. The variance of an instrument is defined as the range, at any given value of the measured quantity, of maximum variation in indication which may be exhibited by the instrument when the measured quantity aperiodically increases and decreases to the same value, after every part of the mechanism has come to rest, and when no change has taken place in the environment.

The *specific variance* or *variancy* is the ratio of this range of indication to the value of the measured quantity.

As far as the writer has been able to determine, the hysteresis loop, and its

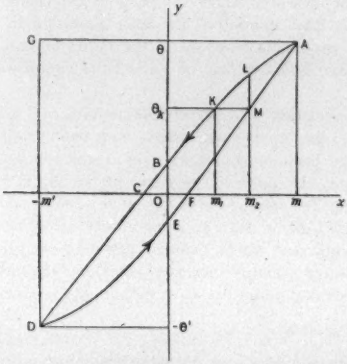


Figure 2.

remarkable constancy for any instrument in a proper working state, is a discovery of the first order made by Schlink. If the hysteresis loop were not constant it would be of no value. Any want of constancy is an infallible indication of injury to the mechanism, of some obscure defect or of faulty workmanship which may be so insignificant as otherwise to escape observation. This has been repeatedly confirmed by the writer. Since the important phenomenon of instrumental variance does not appear to have received attention in precision work with indicating instruments, it will be dealt with in some detail.

Figure 2 represents the hysteresis loop of an instrument with a generalized range of positive and negative values of indication θ plotted along the y-axis and the corresponding values of the measured quantity m plotted along the x-axis.

According to Schlink (1919 a) the area DEAG is proportional to the energy required to produce the deflection of the indicator from the value $-\theta'$ to the value $+\theta$ when the measured quantity changes from $-m'$ to $+m$.

Certain measuring instruments, such as electrical instruments, require in addition a continual supply of energy to maintain the deflection of the

indicator. This is not the case with electrostatic instruments, however, and the energy under discussion is the amount required to cause the movement of the mechanism from one position to another of equilibrium between the force of restitution within the instrument and the force derived from the measured quantity which activates the mechanism.

The area DBAG is proportional to the energy returned by the force of restitution to the agency producing the original deflection upon return of the measured quantity to the value $-m'$. The difference of these areas, which is the area of the loop, is proportional to the energy dissipated in the mechanism.

For any value m_2 of the measured quantity the intercept LM of the ordinate through m_2 is, in accordance with the definition, the instrumental variance at the value m_2 of the measured quantity. The specific term or variancy is LM/m_2.

KM, which is equal to m_2-m_1, is the range of values of the measured quantity for a given deflection θ_k.

At the scale zero, OE and OB are respectively remanent deflections or *sets* when the measured quantity changes from $-m'$ to zero and from m to zero. Their sum BE is the variance at zero reading where the variancy is infinite. OC and OF are respectively residual values of the measured quantity when the deflection changes from θ to zero and from $-\theta'$ to zero.

The efficiency of resilience or the resiliency of the instrument over the cycle is area DBAG/area DEAG.

Expressed analytically, if

$$x=\phi_1(y) \quad \text{and} \quad x=\phi_2(y)$$

are the equations to the increasing and decreasing branches of the loop, the energy dissipated is given by

$$E=\int[\phi_1(y)-\phi_2(y)]dy,$$

and the resiliency by

$$\eta=\int\phi_2(y)\,dy\Big/\int\phi_1(y)\,dy.$$

The term resilience is descriptive of the agency within the mechanism, usually a purposely designed elastic element, which returns the indicator to zero when the measured quantity ceases to act. But whether the restoring agency is gravity or any other force, as that supplied by the hand when operating such instruments as theodolites, polarimeters or calliper micrometers, the variance is manifest and the resiliency must always be less than unity.

In a perfect instrument the resiliency would be unity, and the two branches of the loop would collapse into a single indication/measured-quantity function. Every indication taken at random would be reproducible, and, moreover, the sensitivity at any indication would be single-valued and could be increased to infinitely great values if energy changes were continuous in nature. It is obvious that no such instrument is possible.

The hysteresis loops of various instruments have been determined and described by Schlink (1918, 19b) and by Pollard (1938).

The evaluation of the hysteresis loss in instruments has been discussed by Schlink (1919a).

§6. THE CAUSES OF INSTRUMENTAL VARIANCE

The chief source of variance in instruments is back-lash due to looseness of fit between the mating surfaces of the couplings, which results in relative shifts of component parts of the mechanism without affecting the indicator. Energy is expended in overcoming the friction and the restoring force and, in accelerating the moving parts, it is dissipated by molar and molecular friction. Molar friction depends upon the physical and chemical state of the surfaces and may vary with time.

With those instruments in which spring members are used to restore the indicator to its zero position a source of variance exists in the imperfections of elasticity of the materials used in the construction of such members.

In diaphragm instruments, such as aneroid capsules, the greater part of the variance is due to elastic hysteresis and other elastic irreversible phenomena.

The resistant components of the mechanism not intended to act as springs nevertheless yield under the impressed forces and frequently exhibit marked hysteresis, which has not been purposely reduced to a minimum by special treatment of the materials of construction, and the effects of this and other elastic imperfections may become amplified at the indicator.

A rod or wire, if suddenly stretched by terminal traction, is cooled and, when the traction is suddenly released, is warmed. Similarly a twisted rod, if suddenly twisted further, will cool and will evolve heat if untwisted.

Lord Kelvin (Poynting and Thomson, 1902) established expressions for the relation between the change of temperature and the change of the stretch or squeeze of an elastic body, and showed that two moduli of elasticity exist, the static and kinetic, of which the latter must always be greater than the former.

These irreversible temperature changes will contribute to the variance, and the variation of the elastic modulus due to these changes is important in the choice of materials for elastic systems in which the rate of vibration is required to be as constant as possible. Thus it is advantageous for good time-keeping to construct the hair-springs of chronometers or other timepieces out of the nickel-iron alloy known as elinvar, which has a small temperature coefficient of the elastic modulus.

In liquid-column instruments the variance, which is always present, may arise from changes in the surface tension or the curvature of the meniscus from known and unknown causes.

Sensitive hydrometers, for example, graduated to 0·0001 (specific gravity) over a range of about 0·0150 have been reported (Schlink, 1919a) to exhibit variance amounting to 0·0002 or 0·0003 under laboratory conditions.

The sluggishness of sensitive bubble-levels is well known, and these instruments are usually tapped to give more reliable readings. But the tapping of such sensitive instruments may easily upset the seating of the vial in its mounting, and provision should be made for accurate reversal of the bubble for the purpose of obtaining mean readings.

Bubbles are extremely sensitive to very slight changes of local temperature, and rapidly move under thermal disturbances in a way which is difficult to understand. If, for example, the glass of a sensitive vial, indicating two seconds of arc per 0·1-inch division of the scale, be momentarily and very lightly touched

with the finger at one end of the bubble, the bubble will instantly travel several divisions towards the point touched and will slowly return to its former position.

It may be that such surface-tension phenomena, which may become manifest as variance, are subject to the well known relation between changes of surface tension with temperature (Poynting and Thomson, 1902). But as far as the writer knows, no investigations into the transfer of energy which may take place during the movement of liquid columns have been reported.

The importance of hysteresis determination is emphasized by the peculiar behaviour of the sensitive bubble-level, previously mentioned and reported by Starling (1922–23).

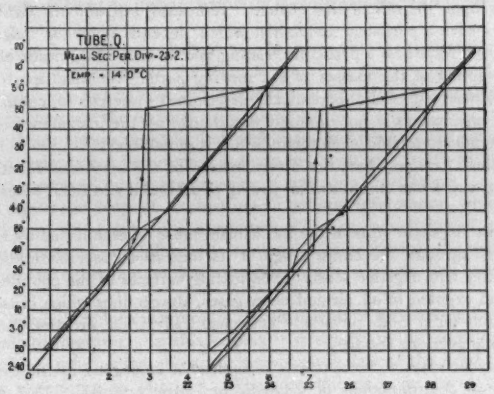


Figure 3.

(*Reproduced, by kind permission, from the* '*Transactions of the Optical Society*'.)

Figure 3 illustrates the displacement of the two ends of the bubble for given tilts of the vial. The loops show that the instrument suffers from marked passiveness which makes it useless. The defect operates while the bubble is moving in one and not in the other direction, and no cause for this behaviour could be discovered.

It should be noted that the branches of the loop cross one another. Since this is not possible, it is evident that no attempt was made to obtain a correct loop, and in the regions of small variance the undetermined and superimposed variance of the tilting instrument upon which the vial was placed has become confused with the variance of the vial.

§7. THE DETERMINATION OF THE HYSTERESIS LOOP

The determination of the hysteresis loop has been fully discussed by Schlink (1919 a), but it is desirable here briefly to state the precautions necessary if constant and reliable loops are to be obtained.

In the first place, since a property peculiar to the mechanism itself is to be

investigated, the environment during the experiment should be constant and similar to that in which the instrument is to be used. Secondly, slow aperiodic motion of the mechanism throughout the experiment is essential. Slow aperiodic motion has the effect of eliminating as disturbing factors: (i) Damping resistances afforded by fluid friction or friction having similar characteristics; (ii) kinetic effects of mechanical friction, such as are directly or indirectly occasioned by oscillatory motion about a shifting rest-point, such that the rest-point becomes indeterminate in the dynamic operation of the instrument; (iii) time-lag engendered by the inertia of the mechanism, a lag of rate of deflection behind the rate of change of the independent variable, and depending primarily upon the rate of that change. Indicating instruments with balanced indicators have a period of oscillation which depends upon the magnitude of the restoring force near a particular reading, and upon the magnitude and nature of the coupling of the masses of the moving parts. So that, apart from the double-valued form of the aperiodically-determined hysteresis function, the indication at any instant is not uniquely determined by the instantaneous value of the independent variable when the indicator is in motion.

For these reasons, then, the determination of the hysteresis function must be carried out in the absence of all jarring or vibration and must be carried out slowly.

Thirdly the *cyclic state* of the mechanism must be established by taking the instrument through the entire range of its indications several times, steadily in a positive and negative sense. The mating surfaces of the couplings, when at rest and exposed to oil, air and other gases, absorb films which cause friction to vary and differ both physically and chemically from surfaces immediately after sliding over one another.

The cyclic state of mechanisms is somewhat analogous to the cyclic state necessary in determinations of magnetic or elastic hysteresis. No doubt the constancy of the elastic hysteresis of components in the mechanism induced by this treatment contributes to the constancy of the loop.

As far as experience goes at the present time, the cyclic state can be induced in all instruments, with the exception of those which exhibit drift phenomena, or in which uncompensated back-lash exists, and which, therefore, require special treatment.

Drift is a time effect characterized by an asymptotic approach of the indicator to the rest-position. It is due mainly to elastic after-effects in the stressed members of the mechanism, or to the viscosity of lubricating oils which take time to be squeezed out of the bearing surfaces and which, consequently, take time to settle into their rest positions. Drift is also due to viscosity of liquids in liquid-column instruments. A hydrometer designed to determine the density of glycerine suffers from drift in an unparalleled degree. The drift phenomena of diaphragm mechanisms, such as aneroid capsules, is well known, and is due to the overstrain in the metal and cold-working effects produced by pressing operations during manufacture.

The phenomenon of elastic-overstrain is well illustrated by the recovery of the sharpness of razor blades after a period of rest.

A valuable summary of early researches on elastic-overstrain is to be found in Todhunter and Pearson's work entitled *A History of Elasticity and Strength of Materials*.

An example of an instrument in which an uncompensated back-lash exists is one in which a pinion integral with the indicator meshes with a rack, without a hair-spring to maintain one-sided contact between the flanks of the teeth. This defect is a fault of design.

After the cyclic state has been induced, the instrument must at once be taken through its loop, but since there is a "set" at the zero value of the measured quantity, the path of the indicator will start at some point within the loop and subsequently will join the upward branch of the loop at some value of the measured quantity greater than zero.

On return there will be a residual "set" at zero value of the measured quantity indicating incompleteness of the loop, as shown in the two loops of figure 5. Consequently, the indicator should be able to pass the zero in the negative direction in order to complete the residual "set" on return.

This is not always possible, but with many instruments the agency of the measured quantity can be reversed in sense, and with such instruments the scale should be continued on the negative side of zero for a short length.

The hysteresis loop of an instrument affords information which can be obtained in no other way: it is a measure of the reproducibility of indications, and a guarantee of reliability. If the loop is not constant, the instrument is unreliable, and either in want of repair or has not been constructed with sufficient care.

The form of the loop is typical of certain classes of mechanism, but generally is difficult to interpret. In this respect it is similar to the electrocardiogram, which affords information about the pathological state of the heart, but which requires experience to interpret. The loop affords a means of comparison of the performance of members of a group of instruments all of the same type of design and construction and, when determined in "trial" mechanisms, is of the greatest value during the development of a design. In many cases the best type of mechanism for any specific purpose can only be selected by means of the hysteresis loop. Thus the best mechanism for the fine adjustment of the microscope to be used for micrometric purposes, if a choice is to be made between the lever, cam or screw type, is at once apparent by comparison of the hysteresis loops.

In order that a reliable quantitative comparison of similar instruments with different scale-ranges may be made, it is necessary to reduce the ranges to a common basis. This reduction can be made if the relation between scale range and loop area is known. Since the minor hysteresis loops of an instrument taken over fractions of the maximum range are bounded by the major loop of maximum range, there should be no difficulty in evaluating the factor of reduction.

The work of Rowett (1913) on torsional hysteresis may be usefully consulted in relation to the reduction factor of the hysteresis loops of instruments employing torsionally-stressed elastic elements.

§8. DAMPING AND RAPID RESPONSE

In addition to the importance of reducing the instrumental variance to a minimum, it is important to reduce the period of movement of the indicator to its rest-position and the oscillations about it—in other words to increase the rapidity of response.

In an ideal instrument the indicator will instantly respond without lag to a change in the measured quantity and come to its rest-position without oscillation. This ideal is nearly realized with cathode-ray oscillographs in which the electron beam, for most purposes, may be assumed to have no mass.

In mechanisms, however, the inertia of the moving system causes the indicator to lag behind the variations of the measured quantity and to oscillate about the rest-position, unless suitable damping is provided.

The whole matter of damping is complicated by friction, which must necessarily also be taken into account.

A great deal of literature exists on this subject, but useful reference may be made to the work of Drysdale and Jolley (1924) and to the paper by Draper and Bentley (1940).

§9. THE ERROR EQUATION AND ITS APPLICATION

As previously stated, functional design should be directed to the ultimate production of a functional model in which the permissible errors of indication lie within specified limits.

Generally speaking, the variance cannot be computed, but can only be determined experimentally after the functional model is constructed. It can, however, be reduced to a minimum by correct design, which will be discussed later.

There generally are, however, certain parameters, such as the dimensions of the components of the instrument mechanism, which affect the readings.

The effect of small deviations of these parameters, from their nominal values, on the indications of the instrument should then be considered in the preliminary survey of the projected design. Let m, the measured quantity, be a function of the independent variable elements or parameters $p, q, r, \ldots$. Then

$$m = f(p, q, r, \ldots),$$

and if $\delta p, \delta q, \delta r, \ldots$ are small first-order variations, their combined effect upon m will produce a variation δm given by Taylor's expansion of the function,

$$\delta m = \frac{\partial m}{\partial p}\delta p + \frac{\partial m}{\partial q}\delta q + \frac{\partial m}{\partial r}\delta r + \ldots.$$

The sign of the terms on the right-hand side of the equation will depend on the signs of δp, δq, δr, etc., together with those of $\frac{\partial m}{\partial p}$, $\frac{\partial m}{\partial q}$, $\frac{\partial m}{\partial r}$, etc.

This equation is known as the *error equation* of the instrument and is universal.

If, now, the probable error of the variations of m, p, q, r, etc., are C_m, C_p, C_q, C_r, etc., then

$$C_m^2 = C_p^2\left(\frac{\partial m}{\partial p}\right)^2 + C_q^2\left(\frac{\partial m}{\partial q}\right)^2 + C_r^2\left(\frac{\partial m}{\partial r}\right)^2 + \ldots.$$

The extreme error of m will occur when the terms in the equation for δm are all of the same sign. The relative importance of the variations of p, q, r, etc., on the total error of the indication of m will in each case be proportional to

$$\left(\frac{\partial m}{\partial p}\right)^2 (\delta p)^2, \text{ etc.}$$

A simple example of an early form of the spherometer may be taken to illustrate the application of the error equation. This instrument, for which Andrew Ross was awarded the silver medal of the Society of Arts in 1841, consisted of a ring, the edge of which has a toroidal curvature of known lesser principal radius and contacts the spherical surface of unknown radius ρ along a small circle. The depth h of the spherical segment thus defined by the ring is measured by the contact at the pole of the segment of the end of a micrometer screw constituting the "scale" of the instrument.

In 1892, Nelson introduced his supposed improvement by reducing the curved edge of the ring to a circular knife-edge of radius r.

The indication/measured-quantity equation for this instrument is

$$\rho = \tfrac{1}{2}\left(\frac{r^2}{h} + h\right).$$

Application of the "error equation" gives

$$\delta\rho = \frac{r}{h}\delta r - \tfrac{1}{2}\left(\frac{r^2}{h^2} - 1\right)\delta h,$$

$$= \delta R + \delta H.$$

So that, in this case, the error $\delta\rho$ of the measured quantity is the sum of two partial errors δR and δH, of which the first δR, or the error due to the ring, is unalterable and is always present in every measurement.

Since the manufactured ring will depart from geometrically circular form and its edge will suffer from imperfections, it will contact the spherical surface along a small circle of radius r with a mean error δr, which cannot be determined until preliminary measurements are made on "trial" rings with spherical surfaces of known curvature. This having been done, it is obviously useless, and certainly uneconomic from the manufacturing point of view, to construct a scale so finely divided that the partial error δH due to it is incomparably smaller than that due to the ring, but the question at once arises, how small must the partial error due to the scale be made ?

Evidently the magnitude of the smallest scale interval, δh, must be such that the mean $\delta\rho$ of the measured quantity is not sensibly increased by the partial error δH.

Since $$\delta\rho^2 = \delta R^2 + \delta H^2,$$

then in order to satisfy this condition δH^2 must not differ from δR^2 by more than a small quantity of the second order.

Consequently δH should not be greater than $0{\cdot}1\ \delta R$, and from this relation δh may be obtained for any given value of ρ and r.

If it is found that δR is too great, the design of the ring must be reconsidered so as to reduce the mean error δr. δh having been determined, a "trial" micrometer screw should be constructed and its loop determined. If the variance

is greater than δh at any reading when contacting spherical surfaces, the design of the screw-spindle coupling must be reconsidered. The investigation is not yet complete, since the effect of eccentricity and want of squareness of the micrometer spindle to the plane of the small circle should be investigated in a similar manner before the manufacturing limits on the component parts of the instrument can be specified.

§ 10. THE REDUCTION OF INSTRUMENTAL VARIANCE

As previously said, the main object of design is to reduce the variance to a minimum or to reduce its value to lie within specified limits at every point of the indication range.

The specified limits will usually reduce to a single value determined by the dimension of the smallest scale reading or "step".

Adjustments do not alter the variance, except to a second order of small quantities.

The variance of a single reading may be reduced by suitably applied vibration or by special manipulation of the instrument.

Vibration causes the couplings to settle into positions of repose under the system of forces acting upon the kinematic links or components, and no doubt also reduces the elastic hysteresis and other molecular causes of variance.

Consequently there is some more or less definite position within the loop towards which the indicator tends to move, in repeated measures, under the influence of the vibrating agency. It is evident that the amplitude of the applied vibrations should gradually decrease to zero before a reading is taken. Electric buzzers attached to the instrument do not comply with this condition, but are preferable to the generally used method of tapping, which may injure delicate mechanisms, such as bubble levels. It would appear that a part of the energy dissipated within the mechanism is supplied by the vibrating agent, and the variance is thereby reduced. The effect of vibration properly applied is astonishing in some cases. Many devices have been described in which the friction of the moving parts is reduced, with a corresponding reduction of the variance. In the Cuttriss modification of Lord Kelvin's syphon recorder, a small piece of iron is attached to the recording pen, which is made to vibrate against the moving surface of the paper by pulsations of an adjacent magnetic field. In the telautograph, a vibratory movement of the pen is maintained by harmonics superimposed on the sending current. Again, in some watt-hour meters, in which the spin of the rotor registers the watts consumed per hour, the brushes which lead the current into the rotor are made to oscillate back and forth, so that the energy lost in friction is, in part, returned.

An ingenious means for effectively reducing frictional loss in a motor-meter has been devised and described by S. Evershed (1899-1900).

The reduction of the variance of a single reading by manipulation is effected by causing the indicator to approach its rest-position from a lower or higher reading. By this means the indicator tends to move along one branch of the hysteresis loop, and consequently tends to rest at the same rest-point, with repeated readings taken in the same direction. This method is generally used when observations are made with theodolites, but it is evident from what has

been said that the best results can only be obtained by this method when the mechanism is in the cyclic state.

The variance throughout the range of indications can only be reduced by correct design and simplicity of the mechanism and good quality of workmanship.

Since a source of mechanical variance is friction and back-lash, the reduction may be effected by reducing the friction-displacement factor both in translatory and rotatory couplings.

An example of the effect of the reduction of this factor upon the hysteresis loop may be seen in figure 4, which shows two loops of a symmetrical slit mechanism

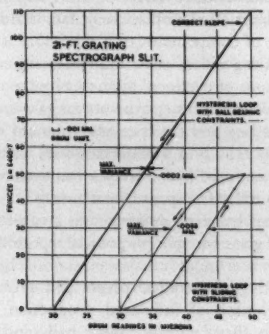


Figure 4.

(Reproduced, by kind permission, from the 'Journal of Scientific Instruments'.)

for a 21-ft. grating-spectrograph previously reported by Pollard (1938). The jaws of the slit are fixed by kinematic constraints to carriages, each of which contacts at five points on an optically worked steel slide-way. The main lever through which the movements of the micrometer head are communicated to the sliding carriages is massive in order to reduce elastic bending to an inappreciable amount. It was found, however, that the hysteresis loop was too large, and is shown in figure 4 with a variance of 0·0055 mm.

One division of the micrometer head is required to represent a movement of the jaws through a distance of 1 μ relative to one another, so that, as the loop shows, for the particular range of the micrometer readings within which it was determined the maximum uncertainty of random setting of the jaws to the indicated amount was $5\frac{1}{2}$ times as great as a step on the micrometer head, and the instrument was untrustworthy. This large variance was obviously due to the excessive friction at the sliding constraints, and the shape of the loop suggests a predominance of elastic hysteresis due to the imperceptible bending of the lever.

This large variance was at once reduced, as the second loop shows, to a maximum of 0·2 μ, when the design was altered and rolling balls were substituted for the sliding constraints.

Since the maximum variance is now one-fifth of the indicated step, the design may be considered satisfactory and the instrument trustworthy, provided the nearest drum-reading only is taken and not estimates of subdivisions of step division, for which the design is not intended.

This instrument has been in continuous use at the Imperial College, South Kensington, for several years, and is just as trustworthy now as when it was first constructed.

That part of the variance due to elastic imperfections obviously may be reduced by increasing the rigidity of the effective components of the mechanism as well as by the reduction of the friction.

The reduction of the friction-displacement factor and of frictional moments, which cause the flexure of components, to a minimum is readily effected by the introduction of ball-bearing slides, ball-rotatory bearings or jewel pivots.

Remarkable reductions of frictional moments by the use of properly designed jewel-pivots and miniature ball-bearing pivots may be achieved.

The permissible loading and the frictional moment of jewel pivots has been investigated by Stott (1931), and a dissertation on materials suitable for these couplings, especially when used in compasses, has been published by Lieut. R. J. Bray (1917) of the Admiralty Compass Department.

Recently special glass has been developed as a substitute for the more costly jewel, and a new hard glass for this purpose is reported by McCune and Goss (1942).

Miniature ball-bearings for conical pivots and cylindrical arbors, of Swiss manufacture, are a triumph of industrial production. The smallest conical pivot bearing made has three minute rolling balls and an overall diameter of 1·5 mm. The races of these miniature couplings are pressed to precise shape and dimensions from sheet steel, and the pivot frictional moment is said to be less than that of a good jewel.

No improvement in the manufacture of knife-edge bearings has been reported within recent years. These couplings are usually made of hardened steel for heavy and of agate for light loads. Recently stellite has been used in place of agate for analytical balances, but no report on the behaviour of this material has been published so far as the writer is aware. Agate is a crypto-crystalline variety of quartz, and the colours of agates are due to the presence of organic matter or traces of iron, manganese or titanium contained in the water from which the stone is deposited, usually in visible layers. Though a hard substance, which takes an excellent polish, agate usually contains minute inclusions and cavities. Consequently it is practically impossible to obtain knife-edges of regular form when the edge is due to the intersection of two facet planes inclined at about 120°, and is not otherwise purposely formed with a radius.

An optical method for obtaining the contour of knife-edges, by means of interference fringes, has been described by Pollard (1937, 38). By this method, the rough and fractured character of the best knife-edges is at once apparent under magnifications of about 350 diameters. By this method also the almost perfect geometric straight edge, of microscopic sharpness, formed by the intersection of facet planes, inclined at about 117°, of carborundum crystals can easily be demonstrated, but it is doubtful whether such sharp edges are practical to use

as knife-edges on account of the high intensity of pressure which would be invoked on the bearing plane by small loads.

Knife-edge bearings are at present constructed with facets and bearing planes which are anything but optical flats, and the object of knife-edge bearings so far as distribution of load along the edge is concerned is thereby defeated. Moreover, the variable contact along the edge during oscillation introduces a source of variance which is apparent in sensitive balances.

Improvements in the manufacture of knife-edge bearings are required to be directed towards the generation not only of optically-flat bearing-surfaces, but also of edges with a cylindrical form which will contact their whole length on an optical flat, and of which the radius of curvature does not exceed a few wave-lengths of light.

Extensive use has been made of thin laminar suspensions to reduce the frictional moment of rotatory couplings.

The use of single thin strip as a coupling to supersede knife-edge bearings, in which the knife-edge has no definite position upon its bearing plane, was described and patented by Artingstall (1862), and Hoghton (1899) patented the use of two thin strips wrapped on parallel cylinders so that they roll together with minimum friction, as if geared with one another.

By forming a hinge with two thin metal-strips crossed approximately at right angles, a rotational coupling is obtained between the two elements to which the strips are fixed. For small motions the axis of rotation is at the point of intersection of the planes of the strips, and the only loss of energy is due to the slight elastic bending of the strips.

The crossed-strip coupling was also patented by Hoghton (1899), and later examined by the French engineer Mesnager (1903). By crossing two pairs of strips, the coupling can be made extremely rigid except for the single degree of freedom of rotation; its application to measuring machines is described by Rolt (1929).

In mechanisms designed on ordinary engineering principles, the variance is mainly due to the looseness of fit between the mating surfaces of turning, sliding and screw pairs. These functional surfaces envelope one another, and in order that there may be relative motion, certain allowances of fit must be specified in order that a distinct clearance may exist between them. If the hollow and solid elements were precisely of the same size, they could not be made to fit, since no surface can be constructed to exact geometric form or with perfect finish.

The limits on the dimensions of the mating surfaces must necessarily be subject to tolerance or errors of workmanship. Consequently, every manufacturing limit must have a specified maximum and minimum value, the difference of which is the tolerance. There will, therefore, be maximum and minimum limits on both the hollow and solid element, which will result in a maximum and minimum allowance of fit between the two elements, which must be a distinct clearance if the two elements are to enjoy relative motion. For fine mechanism or instrument work, experience has shown that a maximum clearance of about 0·001 inch is permissible, whilst the minimum clearance is about a quarter of this amount for a nominal dimension of 0·5 inch. These clearances increase with

increasing size of the nominal dimension and are approximately four times as great with a nominal dimension of 5 inches.

It is thus readily seen that there is a very definite cause of variance in instruments designed on ordinary engineering principles, and it might be concluded that a substantial reduction of the variance could be made by reducing the clearances. But the reduction of the clearances greatly increases the precision and cost of manufacture. In practice it is not possible to reduce the clearance below a certain minimum, which depends upon the geometric perfection as well as upon the physical state of the functional and mating surfaces. The variance of the machine tools which generate the surfaces is more or less reflected in the imperfection of the surfaces they produce. Cylindrical components, for example, with surfaces finished to a high state of perfection in grinding machines, on close inspection are seen not to have geometric circular cross-sections, but rather polygonal forms with a great number of minute flats arranged in complex spiral patterns upon the cylindrical surface. These defects, the cause of which has raised much controversy, as well as the minute scratches which are always produced by the cutting action of the grains of the abrasive, may be removed by lapping with hard materials, such as rouge, in a state of high subdivision and having a higher melting-point than the worked metal. The play of the hard particles between the soft lap, usually made of pitch, and the work raises the temperature of the metal until an almost molecular layer of it melts and flows into a smooth surface, forming what is known as the Beilby layer.

But this process, unless special precautions are taken, results in the production of pseudo-forms. The geometry of these pseudo-forms has been discussed by W. Francis Shaw and Harry Shaw (1932, 32-33), and reference is made to a trilobed cylindrical roller, supplied by the Hoffmann Manufacturing Company, in the *Report* of the National Physical Laboratory for 1919.

These rollers are lapped between precisely parallel plates, and though the diameters of this particular roller were found to vary by not more than 0·00001 or 0·00002 inch when revolved in a vee block under an indicator, it showed variations due to the pseudo-form amounting to 0·001 inch.

In recent years much attention has been given to the perfection of mating surfaces, since it has been demonstrated that wear decreases with increasing smoothness. Many extremely sensitive instruments have been constructed to record the roughness of a surface and methods devised for quantitatively specifying the degree of smoothness. Most of these instruments use a fine stylus, and the vertical movements of the stylus as it traverses a selected part of the surface is highly amplified in a record. A study of this method has been published by the Research Department of Messrs. Taylor, Taylor & Hobson, Ltd. (1944).

When the clearances are reduced to very small amounts on highly finished works, these imperfections may introduce unforeseen interferences of the metal surfaces which will, when the elements are in relative motion, result in variable friction and will seriously affect the variance.

A striking example of this is to be found in tests carried out by Rannie and Dennis (1934) of the Geodetic Survey of Canada, Department of the Interior, Ottawa, on several Wild portable theodolites. In these superbly manufactured instruments, the horizontal and vertical axes are the usual closed turning pairs

with clearances reduced to a minimum, and the functional surfaces of the dead parallel cylindrical elements are finished to mirror perfection. These instruments are guaranteed to read angles within an error of ± 1 second of arc, which is certainly a remarkable claim on such small theodolites. But the tests revealed systematic angular errors of ± 2 to ± 4 seconds of arc. The cause of these errors, which would have been immediately revealed by the hysteresis loops, was traced, after much tedious work, to the form of the cylindrical axes. During rotation there was a periodic tightening corresponding to portions of the mating surfaces, where the departures from true geometric form reduced the clearances to weak interferences. This produced a strain displacement between the verniers and circles. On modifying the design to conform with kinematical principles, as far as the construction would permit, these systematic angular errors vanished.

A design for theodolites, based on kinematical principles which would overcome these difficulties without the necessity of imposing manfacturing limits, has been suggested by Pollard (1935).

A great deal of literature exists on engineering allowances of fit, and many attempts have been made to develop an elastic theory of the interferences necessary for different grades of force-fits. But the unknown variable surface-hardness of manufactured components and the various imperfections, some of which have been referred to above, render these researches practically useless, and reliance, in the present state of knowledge and methods of manufacture, can only be placed upon the average results of extensive practice. These uncontrollable variables which arise in couplings, designed on ordinary engineering principles, are analogous to those which reduce the value of exacting applications of elastic theory to the deformations of simple components in heavy mechanisms. The cross-sectional areas and forms of beams and girders, for example, differ erratically from the assumed values on account of the extensive scaling of the hot metal during the rolling processes, and the elastic constants of the metal can only be assumed to have approximate values. Designers may therefore be assured that graphical methods of solving these problems are sufficiently accurate for general purposes.

But graphical methods are limited and, like the more exact analytical theory of elasticity, can only be applied to certain simple three-dimensional forms, which occur only occasionally in instrument mechanisms. Simple graphical methods have, however, been devised (Pollard, 1921) for dealing with any form of cross-section in beam and girder problems.

These problems do not frequently arise in instrument design, but when it is necessary to investigate the distribution of the principal stresses in complex forms of components, or to reduce undesirable localization of stress exceeding 50 per cent of the elastic limit, the methods of photo-elasticity can be used. The principles and practice of photo-elasticity are fully dealt with in a treatise on the subject by E. G. Coker and L. N. G. Filon (1931), and more recently by Frocht (1941). The former comprehensive work is limited to uni-planar photo-elasticity, but, on account of a remarkable property of bakelite, which is an excellent material, with a high stress-optical coefficient, for photo-elasticity models, photo-elasticity methods have been extended to three-dimensional analysis by Hetényi (1938–39).

Bakelite when loaded at an elevated temperature retains the applied stress by polymerization after the load is removed at room temperature. The model can then be sectioned and the distribution of stress investigated.

During manufacture the dimensions of the components are controlled by means of limit gauges, and the general principles of gauging in the workshop are briefly described by Rolt (1929). When the functional model is constructed, however, each component serves as the gauge for the mating component and the necessary fit is made by the workman. The sizes of the model then become the basic dimensions on which the limits can be established for interchangeable manufacture.

For very small-quantity production the same method of fitting is usually applied, but for large-quantity production this method is highly uneconomic and depends upon the skill and judgment of workmen. It is then necessary to establish a limit system which must be maintained by the necessary limit gauges. This is a costly undertaking and requires great care and detailed study.

The designer before embarking on an undertaking of this magnitude should make a study of interchangeable manufacture, the principles of which are ably described in an authoritative work by Earle Buckingham (1921), who was at one time Engineer to the Pratt and Whitney Company of America.

Buckingham particularly warns the designer that the condition of "compound tolerances" must be avoided since, when the variation of a dimension causes a variation of another dimension, the interpretation of the drawing will be ambiguous unless particular care is taken in indicating the dimensions, and more than one unique component will be produced. Consequently, when such mating components are selected at random, they will not fit in the manner desired.

Moreover, it is important, when draughting a design, to select basal surfaces from which the dimensions are measured in the shop, otherwise unnecessarily small tolerances may be imposed.

Since no design of a mechanism can be considered to be complete unless provision is made for checking special component measurements, which cannot always be checked by the usual workshop measuring tools, it must be accompanied, when necessary, by the design of special measuring devices. The designer should therefore be fully acquainted with the common measuring methods and instruments used in the workshop, and cannot do better than study the excellent work of Rolt (1929) on this subject.

From what has been said it is clear that that part of variance which arises from the looseness of fit of mating surfaces in the couplings of instrument mechanisms can only be substantially reduced by eliminating the clearances altogether. This may be achieved by kinematic design, which for this and other reasons is preferable to ordinary engineering design.

It is generally believed that Robert Willis, Jacksonian Professor of Natural and Experimental Philosophy in the University of Cambridge, was the first to apply kinematical principles to the design of mechanisms. He published a treatise on the *Principles of Mechanism* in 1841, the year the illustrious Kelvin entered St. Peter's College as a student at the age of 17, but though he did not mention these principles in his book, Kelvin subsequently acknowledged his indebtedness to Willis for a knowledge of them. There are reasons for thinking,

however, that Willis was in fact anticipated, at the commencement of the 19th century, by Sir John Barton, who ruled the famous diffraction buttons, known by his name, with an engine designed on kinematical principles.

Only two examples of kinematic design were published by Willis as far as the writer is aware. One is the design of his model illustrating crank motion and the other a design for a lathe tool-holder, which is described and illustrated on p. 996 of the second volume, by Charles Holtzapffel, of the extensive work, entitled *Turning and Mechanical Manipulation* (Holtzapffel, C. and Holtzapffel, J. J., 1866-97).

Kinematic design is based on the obvious and simple theorem stated by Maxwell, i.e. a rigid body has six degrees of freedom.

If co-ordinate axes are fixed in one body or link of a kinematic chain, any displacement of another link relative to the first may be resolved into three displacements parallel to the axes and three rotations about them.

To fix the one link to the other, these possible motions about the axes must be prevented by a system of forces properly applied between the two elements of the coupling. Thus, if six points—consisting, say, of convex bosses—properly chosen on the barrel and stock of a rifle be made to rest on six surfaces of a rigid stand, the rifle will be fixed when placed upon the stand and may be replaced any number of times in precisely the same position for the purpose of testing the accuracy of fire.

Moreover, if the contacting bosses are the ends of adjusting screws, which can be locked in position, the rifle can be delicately adjusted with reference to the axes of the shooting range. Such adjustments would otherwise require elaborate sliding and rotatory pairs, with their attendant variance. This possibility of introducing in a simple manner adjustments between the two elements of a kinematic coupling eliminates precision in manufacture and the necessity of replacing worn parts.

If, now, one point only of the one element makes contact with a surface of the other, represented, say, by the yz-plane, a translation in one direction parallel to x is prevented and the elements of the couplings have a single constraint and five degrees of freedom relative to one another. It is easy to see that every additional and properly chosen constraint produced by the contact of a point or convex surface formed on the one element with a surface formed on the other will reduce the relative motion of the two elements of the couplings by one degree of freedom.

It will thus require five constraints properly chosen to give the single degree of freedom of the lower pairs, and six constraints to fix one link to the other.

A single constraint will consist of constraining surfaces, i.e. a curved surface contacting another surface which may be regarded as the *bearing surface* of the constraint. In practice it is convenient to use ordinary steel balls as the curved surfaces, and to construct the bearing surfaces in the form of planes.

The proportions of these elements which it is convenient to adopt in the practical construction of ball-bearing slide-races and the derived kinematic constraints are discussed and tabulated by Pollard (1934).

Since there are no clearances with these constraints no limits are necessary, and the exact location of the bearing surfaces is unimportant. Consequently,

workmanship can be concentrated on the production of true geometric form and high finish of the constraining surfaces without the complication of exacting dimensions. This is a very different and much simplified manufacturing proposition.

The kinematic design of couplings thus offers far-reaching economic manufacturing advantages as compared with ordinary engineering design, and considerable advantages to the user who understands and desires to obtain the best performance of his instrument. But the outstanding advantage is that the variance of instruments is reduced to a minimum.

The two hysteresis loops of figure 5 will illustrate the superiority of kinematic over ordinary engineering design. The instruments concerned are precisely similar reading microscopes in so far as their kinematical elements are concerned. They consisted of a microscope attached to a member sliding in the framework

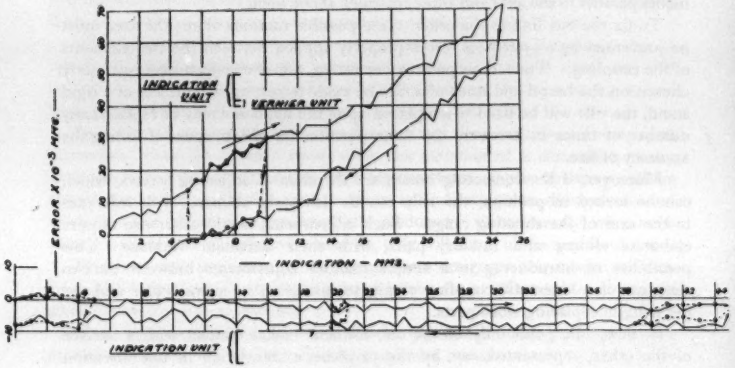


Figure 5.

(*Reproduced by courtesy of Adam Hilger, Ltd.*)

coupled to a micrometer screw, the nut of which is fixed to the framework, so that the simple closed kinematic chain of both mechanisms is : framework—micrometer screw—thrust—sliding member—framework, the framework being the fixed link. But the couplings of the instrument constructed by the Société Genevoise, whose loop is shown in the upper part of figure 5, are designed on ordinary engineering principles, screw and ordinary nut, thrust collar and dovetail slide, whereas the lower loop was obtained from an instrument manufactured by the Cambridge Instrument Company, in which all the couplings are kinematically designed.

The upper loop shows a maximum variance $2\frac{1}{2}$ times as great as the indication step of 0·01 mm. and $12\frac{1}{2}$ times as great as the vernier step, which consequently is useless and misleading.

The lower loop shows a maximum variance less than the indication step of 0·01 mm.

An examination of the first instrument indicated that the mechanism was manufactured with the greatest care and precision, to minimize back-lash, which was unnecessary with the second.

As previously stated, if one of the six constraints is removed, the coupling will enjoy one degree of freedom, which may be a translation, a rotation or a degree of freedom of the most general character, namely, a velocity-twist. Thus there are three types of coupling with one degree of freedom corresponding to the sliding and turning pairs and, when the pitch of the velocity-twist is constant, to the screw pair.

Every constraint removed introduces a corresponding degree of freedom, and thus there are six classes of couplings with 0 to 5 degrees of freedom. In each class there are possible types of couplings depending upon the type of freedom, and each type contains sub-types which depend upon the number of different bearing surfaces, which constitute the single coupling.

In the first class, generally termed "geometric clamps", there is no freedom, and, therefore, there can only be a single type of coupling with at least four subtypes with six, five, four and three different bearing surfaces. In the second class of one degree of freedom there are, as we have seen, three types of coupling. In the third class with two degrees of freedom, there is a type of coupling with two rotations, a type with one translation and one rotation, but the type with two translations cannot exist, for if one element of the coupling is restrained to move parallel to one axis only, there can be no translation perpendicular to this axis. Thus in the third class there are only two types. In the fourth class there are three types instead of four, in the fifth class two instead of three, and in the sixth class one instead of two, for the same reason. All these classes have their subtypes. With six degrees of freedom there are no point constraints. But hydrometers and specific-gravity balls illustrate this case in which the supporting and fixed link is a fluid exerting surface constraint. The indicator of the former is a liquid meniscus and of the latter the instrument itself. But neither instrument is free from variance. Though the gravity ball is theoretically a null-reading instrument, in practice differences of small motions are observed.

There are thus twelve possible types of practical kinematic couplings and a number of subtypes.

When designing kinematic couplings it is necessary to remember that the bearing surfaces should be so chosen that, if one of them is removed, the direction in which the corresponding contacting point moves in virtue of the remaining constraints should be in a direction normal to the tangent plane at the bearing. This condition implies that of the normals to the tangent planes at the bearings no two coincide; no three are in one plane and either meet in a point or are parallel; no four are in one plane, or meet in a point, or are parallel or, more generally, belong to the same system of generators of an hyperboloid of one sheet. The conditions for five and for six normals are more complex (Maxwell, 1890).

A proposition of the greatest importance immediately flows from the kinematical theorem. It is this: If the constraint between two members exceeds six in number, the members will suffer unnecessary elastic deformation. For if the two members are relatively fixed by six constraints, it will require forcible

contact or the most exacting workmanship at a given temperature to enable a seventh non-adjustable constraint to be introduced.

As the two contacting surfaces of a constraint can move apart, a force or forces, referred to as a "force-closure", must be applied to the two elements of the coupling to maintain contact between the constraining surfaces. Since the position and direction of the constraining forces are defined by the points of contact and the normals to the bearing surfaces, there can be no difficulty in arranging the constraints and force-closure to form a system in static equilibrium, but the point or points of application of the force-closure should not be confused with the constraints, which cannot exceed six in number.

In every case it is important to remember that since the constraints are definitely located, and all precision with regard to dimensions is unnecessary, the force-closure must be elastic and capable of exerting the necessary closing force by the elastic flexure of the force-closing component or components. Otherwise the coupling will be unnecessarily strained, and the avoidance of unnecessary strain is one of the main advantages of kinematic design. It is unfortunate that in some otherwise excellent designs on the market this very necessary condition has been violated. Kelvin's arrangement of six constraints, usually known as the "hole, plane and slot" arrangement, is useful when, for example, an instrument is required to stand in a fixed position with the three rounded feet contacting the six bearing surfaces by its own weight. In this case the coupling is closed by the force of gravity acting through the centroid of the instrument, which is equal and opposite to the resultant of the system of forces acting at the constraints.

When the constraints are adjustable, the elements of the coupling can be given relative displacements which are sometimes important to define. Usually there is no difficulty in determining these displacements, but any difficulties can at once be resolved by observing the movements of a "trial" coupling before the final design is completed.

In Professor Horace Lamb's book *Higher Mechanics* (Cambridge University Press) much information of use to the designer of kinematic couplings will be found.

The limiting static load which a kinematic constraint, consisting of a spherical surface in contact with a plane, will tolerate without damage is greater than would appear at first sight, otherwise such components as rotating and thrust ball bearings could never have survived in heavy engineering practice.

The elastic phenomenon of contact between a sphere and plane is remarkable, and the mathematical theory of the contact of isotropic elastic bodies was first developed by the genius of Hertz (1881) and afterwards considerably developed and extended by Sundberg (1919).

The simpler theory of Hertz, however, is sufficiently accurate for purposes of design.

By applying the results of this theory to spherical surfaces of diameter D contacting with a plane, the safe load P is equal to SD^2, where S is the safe load for unit diameter of the sphere. The value of S is proportional to the Hertzian hardness, and is a function of the elastic constants of the material of the plane, provided the hardness of the sphere is greater than that of the plane.

The results of research work on the method of determining the Hertzian hardness and the value of S for several materials has been published by Phillips and Pollard (1937), and a summary of the values obtained for S with some materials is given in table 1, where P and D are expressed in pounds weight and inches respectively.

H is the maximum stress in the material corresponding to the safe load, and it greatly exceeds the elastic limit for steels.

Table 1

Material	Treatment	Vickers diamond hardness	S lbs. per sq. inch	H tons per sq. inch	Approximate elastic limit in compression tons per sq. in.
0·9 C. steel	Dead hard	883	348	187	
,,	Super-hardened	1000	1040	270	
High carbon chrome steel (Hoffmann)	Hard, as supplied	821	520	211	0·83 % C. steel may
,, ,,	Drawn at 210° c.	732	640	—	reach 43 tons
,, ,,	Drawn at 315° c. (spring temper)	657	730	247	per sq. inch
,, ,,	Drawn at 480° c.	511	400	—	
,, ,,	Annealed	259	72	111	
Brass	—	108	6·0	37·2	12
Phosphor-bronze	—	151	17·2	57·6	—
Cast iron	—	177	16·0	53·4	20

It will be seen that S has no relation to Vickers diamond hardness numbers, but the most remarkable result is that high carbon chrome steel let down to a blue spring temper is able to tolerate a higher load than when dead hard.

The high value of S for superhardened 0·9 C. steel by the cloud-burst method is also interesting, but rather what might be expected from the surface cold-working the steel receives.

These results also show that cast iron is an unsuitable material for kinematic constraints.

Measurements were not made on the hard alloys, such as stellite, cutanit, wimet, carboloy or norbide (boron carbide), the hardest substance known next to diamond, but no doubt these would yield very high values of S if both constraining surfaces were constructed from such materials.

If it is found that the loading is too great for a "point" constraint to tolerate, the convex surface can be expanded into a "pad" of sufficient area to take the load safely, but in order that this shall be the case, the pad must be fitted to the bearing surface, and the principles of kinematic design are at once violated.

This difficulty, however, can be overcome by the introduction of an extra link in the kinematic chain. Thus, if a shallow washer is introduced between the spherical and plane surface of the constraint, such that the spherical surface is seated upon the edge of the hole in the washer and the flat of the washer contacts with the plane, the compound coupling thus formed will tolerate very much greater loads, and the two elements of the coupling will still enjoy five degrees of freedom.

An example of kinematic design may be given to illustrate the satisfactory manner in which ball-plane constraints will tolerate moderately heavy loads.

Figure 6 gives a diagrammatic view of the Cambridge Instrument Company's spectroheliograph, in use at the Solar Physics Observatory, Cambridge, previously described by Pollard (1933 b).

This instrument, which embodies the well-known Janssen principle of monochromatic photography of the sun's disc, was designed by Sir Horace Darwin for Sir Norman Lockyer. The image of the sun, about 5·5 cm. in diameter, is cast by means of a heliostat and a projecting lens upon the primary slit of a spectroscopic train. The dispersed beam is reflected back parallel to its original direction by means of a mirror, and is brought to a focus on the photographic plate. The large triangular frame A carries the primary slit and spectroscopic train and is constrained kinematically to move with one degree of translation parallel to the plane of the primary slit. The frame rolls on three

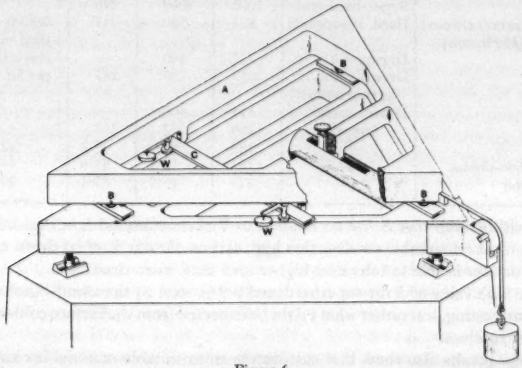


Figure 6.

(*Reproduced, by kind permission, from the 'Proceedings of the Institution of Mechamcal Engineers'.*)

balls B, 1 inch in diameter, loosely located between hardened and lapped parallel steel planes. To limit the motion to one degree of translation two more constraints are necessary, and these are formed by two projections in the frame bearing against the hardened and lapped edges of a steel guide C by means of two rollers and counterpoises W.

To control the rate at which the frame is allowed to move under the action of a suspended weight, a cylinder filled with oil is employed, the movement of the frame causing a plunger to enter the cylinder and expel the oil from it. A valve controls the escape of oil and can be set micrometrically to vary the total time of traverse of the frame, and thus to vary the exposure on the photographic plate, from a few seconds to more than an hour.

The writer was indebted to Mr. Butler, who was in charge of the instrument in 1933, for the following information. He estimated that the total load on the balls is about 4 cwt. After sixteen years of continuous service, slight indentations

developed in the steel planes. The balls were then merely shifted sideways into new positions. Twelve years afterwards, i.e. twenty-eight years after the instrument was erected and had been in continuous use, the plates were reconditioned. The sliding movement is about 6 to 7 cm. in length and has to be very uniform, otherwise a number of dark lines make their appearance across the sun's photographic image. These lines correspond to momentary changes of velocity of travel and indicate variations of exposure on the plate as the slit traverses the sun's image.

These lines made their appearance when the instrument was first used, and the cause of the irregular motion was found to be due to the use of twisted wire cord for the falling weight. The irregularities on the surface of the cord varied the pull as they passed over the groove in the pulley, and caused the oil in the controlling gear to escape at a variable rate. The defect was at once remedied by replacing the cord with a steel tape of uniform thickness and the pulley by one with a cylindrical surface. This will give some idea of the regularity of the motion required and the perfection of the simple mechanism which made it possible.

Kinematic design was first introduced into industry by Lord Kelvin, as evidenced by his elegant and famous instruments manufactured by James White, Optician, of Glasgow, in which firm he became a partner.

Following him, Sir Horace Darwin (Darwin, 1913; Darwin and Mason, 1923), who was influenced by Maxwell, applied kinematic design to many instruments constructed by the Cambridge Instrument Company, established by him in partnership with Dew-Smith in 1881. A notable early example of his kinematic designs is the now well-known rocking microtome. This precision tool, capable of cutting serial sections 0·0001 inch thick and maintaining its accuracy over many years of service, is an outstanding example of accuracy of action, robustness and simplicity of manufacture. Many kinematically designed instruments have been produced in recent years by this company as well as by other firms in this country, such as Messrs. Adam Hilger, Ltd., Taylor, Taylor and Hobson, Ltd., of Leicester, W. G. Pye and Co., of Cambridge, Alfred Herbert, Ltd., of Coventry and Daniel Varney, Ltd., of London. In particular, Messrs. Cooke, Troughton and Simms, Ltd., of York, have recently devised a simple kinematic self-centring mount for theodolite divided glass circles with a single reading point.

This type of design has also found favour with the National Physical Laboratory, and many elegant applications are described in the Reports and by Rolt (1929). Successful designs by individual workers, which appear to be on the increase, are described in the literature, and the application of kinematic design has especially received attention from R. S. Whipple (1921, 20-21), J. E. Sears (1923), J. A. Tomkins (1928), and A. F. C. Pollard (1922, 29, 33 b).

In other countries kinematic design has so far only been applied to a few commercial productions, such as the ball-bearing slide of the fine adjustment of microscopes manufactured by the firm of Ernst Leitz, but descriptions of designs by individual workers appear from time to time in the literature, such, for example, as the adjustable slit mechanism of the large spectrograph of the Laboratoire de Recherches magnéto-optiques at the Office national des Recherches et Inventions (de Gramont, 1928).

§11. THE DESIGN OF SAND-CAST AND MACHINED COMPONENTS

As previously stated, it is not possible in a short *Report* to deal with all the limitations imposed on instrument design by the various workshop processes necessary for the production of component parts, but attention will be directed to the most important points to be observed when designing parts which are to be sand-cast and subsequently machined.

There are, however, three primary desiderata from the manufacturing point of view which should be considered in the design of any mechanism.

Firstly, every component part should be considered from the point of view of practical construction to ensure economic and possible production with the manufacturing tools and plant available in that factory where the mechanism is to be constructed. It is, for example, bad practice to call for screws with Lowenhertz threads when no tools are available for cutting this particular thread form. To make such tools would involve an unnecessary expenditure and waste of time and labour.

Secondly, standard stock components and standard materials such as screws, rod, sheet, drawn or rolled sections, and so forth, which are the products of specialized manufacture, should be used whenever possible.

Thirdly, every endeavour should be made to adhere to national standards, of which British Standards are published by the British Standards Institution, of 28 Victoria Street, Westminster, London, S.W. 1.

Almost any form of casting can be successfully produced, at a cost, when special skill and the results of recent research are available. But with ordinary foundry practice there are a number of points in design which should be observed if clean and sound castings are required.

After the casting is produced it has to be machined, and here again there are several points of design which should be observed if the machining operations are to be carried out without unnecessary trouble and expense.

These details have been discussed by Pollard (1933 a), and it is only necessary here to summarize them as follows:

(1) Avoid complicated shapes; design simple shapes easy to draw from the mould. Use two flasks rather than three.

(1.1) Design two or three simple forms to be bolted together, rather than one troublesome casting.

(1.2) Machined faces should be accessible.

(1.21) Avoid machined faces in enclosed positions.

(1.211) Avoid internal slotting.

(1.22) Allow sufficient landing space for tools in planing, milling, grinding and screwing operations.

(1.23) Avoid parallel machined surfaces at different levels.

(1.24) Avoid cast lugs, bosses or other extensions on machined surfaces. Screw and dowel separate pieces in position.

(1.3) Avoid loose pieces, such as bosses on patterns.

(1.31) Avoid insufficient metal in bosses.

(2) Avoid coring whenever possible.

(2.1) Especially avoid coring on small castings.

(2.2) Avoid large cores with small prints.

(2.3) Place cores at top rather than bottom of mould.

(2.4) Place elongated cores likely to bend vertical in the mould.

(3) Avoid sudden changes in section. Unequal shrinkage and subsequent fracture likely.

(4) Avoid sharp corners and projecting lugs, or localized stress may develop.

(5) Drilling. For accuracy boring preferable, especially if blow-holes are present.

(5.1) Avoid sloping faces to axes of holes at entry and exit of drill.

(5.2) Avoid blind holes. Design for use of boring-bar.

(6) Avoid fine clearances on castings; accidental shift of sand will cause "scabs" and subsequent fouling may result.

(7) Avoid raised letters or figures of small dimensions on surfaces. Replace rather by engraved plates.

(8) Design component so that it may be readily held or secured during machining operations, with the usual chucks, clamps, dead and live centres, etc.

(8.1) In small-quantity production avoid necessity of constructing special securing tackle, by designing additional projections on casting for subsequent removal by machining.

In conclusion, it cannot be too persistently emphasized that any design which, curiously enough, very often commences with unnecessary complexity should be reduced by detailed study to the utmost possible simplicity.

This primary desideratum was recognized by the genius of Leonardo da Vinci (MacCurdy, 1938), who wrote:

"When you wish to produce a result by means of an instrument do not allow yourself to complicate it by introducing many subsidiary parts, but follow the briefest way possible, and do not act as those do who, when they do not know how to express a thing in its own proper vocabulary, proceed by a method of circumlocution and with great prolixity and confusion."

GENERAL REFERENCES

Bouasse, H., 1917. *Construction, Description et Emploi des Appareils de Mesure et d'Observation* (Paris: Librairie Delagrave).

Bray, R. J., 1917. *Pivots and Caps in Compasses* (London: H.M. Stationery Office).

Brunt, D., 1931. *The Combination of Observations* (Cambridge: University Press).

Buckingham, E., 1921. *Principles of Interchangeable Manufacturing* (London: The Machinery Publishing Co.).

Chauvenet, W., 1891. *A Manual of Spherical and Practical Astronomy*, **1** and **2** (Philadelphia: J. B. Lippincott).

Coker, E. G. and Filon, L. N. G., 1931. *A Treatise on Photo-elasticity* (Cambridge: University Press).

Drysdale, C. V. and Jolley, A. C., 1924. *Electrical Measuring Instruments* (London: Ernest Benn).

Frocht, M. M., 1941. *Photoelasticity* (London: Chapman and Hall).

Holtzapffel, C. and Holtzapffel, J. J., 1866-97. *Turning and Mechanical Manipulation*, **1–6** (London: Holtzapffel and Co.).

Lamb, H., 1920. *Higher Mechanics* (Cambridge: University Press).

Pollard, A. F. C., 1929. *The Kinematical Design of Couplings in Instrument Mechanisms* (London: Adam Hilger, Ltd.).

POLLARD, A. F. C., 1933 b. *Proc. Inst. Mech. Engrs.*, **125**, 143.
REULEUX, F., 1876. *The Kinematics of Machinery* (London : Macmillan).
ROLT, F. H., 1929. *Gauges and Fine Measurement*, 1 and 2 (London : Macmillan).

OTHER REFERENCES

AMPÈRE, A. M., 1834. *Essai sur la Philosophie des Sciences* (Paris : Bachelier).
ARAGO, D. F. J., 1853. *C.R. Acad. Sci., Paris*, **36**, 276.
ARAGO, D. F. J., 1859. *Mémoires Scientifiques*, **2**,* 233 (Paris : Gide).
ARTINGSTALL, F. D., 1862. *British Patent* 2469
BABBAGE, C., 1826. *Phil. Trans.*, **116**, Pt. II, 250.
BALL, R. S., 1908. *A Treatise on Spherical Astronomy*, 462 (Cambridge : University Press).
CLERK, H., 1863. *Proc. Roy. Soc.*, **12**, 453.
COKER, E. G. and FILON, L. N. G., 1931. *A Treatise on Photo-elasticity* (Cambridge : University Press).
DARWIN, H., 1913. *Aeronautical J.*, July.
DARWIN, H. and MASON, C. C., 1923. *Dictionary of Applied Physics*, **3**, 445 (London : Macmillan).
DE GRAMONT, A., 1928. *Revue d'Optique*, 7e Année, 68.
DRAPER, C. S. and BENTLEY, G. P., 1940. *Trans. Amer. Soc. Mech. Engrs.*, **62**, 421.
EVERSHED, S., 1899–1900. *J. Inst. Elect. Engrs.*, **29**, 743.
GORE, G., 1879. *Proc. Roy. Soc.*, **30**, 32.
GUILLAUME, C.-É., 1927. *L'Oeuvre du Bureau International des Poids et Mesures*, 167.
HEARSON, T. A., 1896. *Phil. Trans.*, **A, 187**, 15.
HERSCHEL, J. F. W., 1858. *Outlines of Astronomy*, 99 (footnote) (London : Longman, Brown, Green, Longmans and Roberts).
HERTZ, H., 1881. *J. für die reine und angewandte Mathematik*, **92**, 156. (A translation is to be found in *Miscellaneous Papers*, Heinrich Hertz, 1896 : Macmillan).
HETÉNYI, M., 1938. *J. Appl. Mech.*, **5**, A. 149.
HETÉNYI, M., 1939. *J. Appl. Phys.*, **10**, 295.
HOGHTON, H. P., 1899. *British Patent* 11,485.
ISING, G., 1926. *Phil. Mag.*, **1**, 827.
ISING, G., 1931. *Ann. Phys., Lpz.*, **8**, 905.
JEFFREYS, H., 1931. *Scientific Inference* (Cambridge : University Press).
MCCUNE, F. K. and GOSS, J. H., 1942. *Trans. Amer. Instn. Elect. Engrs.*, **61**, 673.
MACCURDY, E., 1938. *The Notebooks of Leonardo da Vinci*, **1**, 71 (London : Jonathan Cape).
MAXWELL, J. C., 1890. . *Scientific Papers*, **2**, 505 (Cambridge : University Press).
MELLOR, J. W., 1916. *Higher Mathematics for Students of Chemistry and Physics*, 514 (London : Longmans, Green and Co.).
MESNAGER, A., 1903. *C.R. Acad. Sci., Paris*, **137**, 908.
MONCRIEFF, R. W., 1943. *Manufacturing Chemist*, **14**, 33.
MONCRIEFF, R. W., 1944. *Manufacturing Chemist*, **15**, 281.
ORNSTEIN, L. S., and MILATZ, J. M. W. 1938. *Physica*, **5**, 971.
PHILLIPS, A. C. and POLLARD, A. F. C., 1937. *Engineering*, **143**, 223, 281, 339 and 398.
POLLARD, A. F. C., 1921. *Machinery*, **19**, 63.
POLLARD, A. F. C., 1922. *J. R. Soc. Arts*, **70**, 769, 783 and 797.
POLLARD, A. F. C., 1933 a. *J. Sci. Instrum.*, **10**, 265.
POLLARD, A. F. C., 1934. *J. Sci. Instrum.*, **11**, 173.
POLLARD, A. F. C., 1935. *Conference of Empire Survey Officers, Report of Proceedings*, 88.
POLLARD, A. F. C., 1937. *J. Soc. Chem. Ind.*, **15**, 315.
POLLARD, A. F. C., 1938. *J. Sci. Instrum.*, **15**, 37.
POYNTING, J. H. and THOMSON, J. J., 1902. *A Text Book of Physics ; Properties of Matter* (London : Charles Griffin and Co.).
RANNIE, J. L. and DENNIS, W. M., 1934. *Canad. J. Res.*, **10**, 347.
RAYLEIGH, Lord, 1893. *Proc. Roy. Soc.*, **53**, 134.

* This constitutes volume 10 of *Oeuvres de Francois Arago*.

ROWETT, F. E., 1913. *Proc. Roy. Soc.*, A, **89**, 528.
SCHLINK, F. J., 1918. *Bureau of Standards Scientific Paper* No. 328.
SCHLINK, F. J., 1919 a. *J. Franklin Inst.*, **187**, 147.
SCHLINK, F. J., 1919 b. *J. Wash. Acad. Sciences*, **9**, 25.
SCHOENEICH, H., 1933. *Schraubensicherungen* (Berlin: Carl Heymanns Verlag.).
SEARS, J. E., 1923. *Dictionary of Applied Physics*, **3**, 570.
SHAW, W. F. and SHAW, H., 1932. *Machinery*, **40**, 777.
SHAW, W. F. and SHAW, H., 1932–33. *Machinery*, **41**, 40 and 349.
STARLING, S. G., 1922–23. *Trans. Opt. Soc.*, **24**, 261.
STOTT, V., 1931. *Coll. Res. Nat. Phys. Lab., Lond.*, **24**, Paper 1.
SUNDBERG, K., 1919. *Teknish Tidskrift, afd. Mekanik*, No. 4.
TAYLOR, TAYLOR and HOBSON, Ltd., 1944. *Report on the Measurement of Surface Finish by Stylus Methods* (Leicester: Taylor, Taylor and Hobson, Ltd.).
TOMKINS, J. A., 1928. *Instrument World*, **1**, 229.
TURNER, L. B., 1937. *J. Instn. Elect. Engrs.*, **81**, 412.
WHIPPLE, R. S., 1921. *Engineering*, **41**, 659.
WHIPPLE, R. S., 1920–21. *Trans. Opt. Soc.*, **22**, 35.
WRIGHTSON, T., 1879. *J. Iron Steel Inst.*, 418.
WRIGHTSON, T., 1880. *J. Iron Steel Inst.*, 11.

THE LIGHTNING DISCHARGE

By J. M. MEEK * and F. R. PERRY,
Research Department, Metropolitan-Vickers Electrical Co., Ltd.

§1. INTRODUCTION

The lightning discharge and its electrical characteristics have been investigated extensively, particularly during recent years, and a fairly clear picture of the development of the discharge process has now been established. This has been made possible largely by the application of high-speed recording techniques involving both photographic and oscillographic apparatus. In the present *Report* a description is given of the results of these investigations, together with a brief summary of knowledge concerning the generation of electric charge within the thundercloud. No claim is made that this *Report* includes a complete survey of the extensive literature concerning lightning. It is intended rather to summarize present knowledge as based on the results of the most recent work, and for that reason little reference is made to many of the early publications on the subject. Because of the incomplete nature of present theories concerning the lightning mechanism, greater emphasis is given in this *Report* to the experimental evidence than to the theoretical treatment of the subject.

§2. THE THUNDERCLOUD

Considerable progress has been made during recent years in our knowledge of the distribution of electricity in thunderclouds as the result of experiments carried out by Simpson with Scrase (1937) and with Robinson (1941). By means of apparatus attached to free balloons, released at intervals during the passage of the thundercloud overhead, records have been obtained of the variation of electrical field in the atmosphere between the ground and the cloud and within the cloud itself. The apparatus included an alti-electrograph, which recorded the variations with time of the atmospheric pressure, from which heights could be determined, and also the sign and approximate magnitude of the vertical electrical field. The latter record is produced on a disk of specially prepared paper on which rests a pair of electrodes, one connected to the instrument case and the other to an aerial about 25 m. long trailing below the balloon. According to the polarity of the field, a stain is produced under one or other of the electrodes, and the width of the stain varies with the external electrical field.

Results of soundings by balloons during two separate thunderstorms are shown in figure 1, together with records of the vertical field observed at the ground and of the charge brought down by rain. The soundings are plotted with a width proportional to the trace on the pole-finding paper, and after calibration give a basis for estimation of the magnitude of the electrical field. The direction of the field is positive where the vertical band is shaded, and is negative where the band is unshaded. A positive field is defined as one in which the force on an electron is

* Now Professor of Electrical Engineering, Liverpool University.

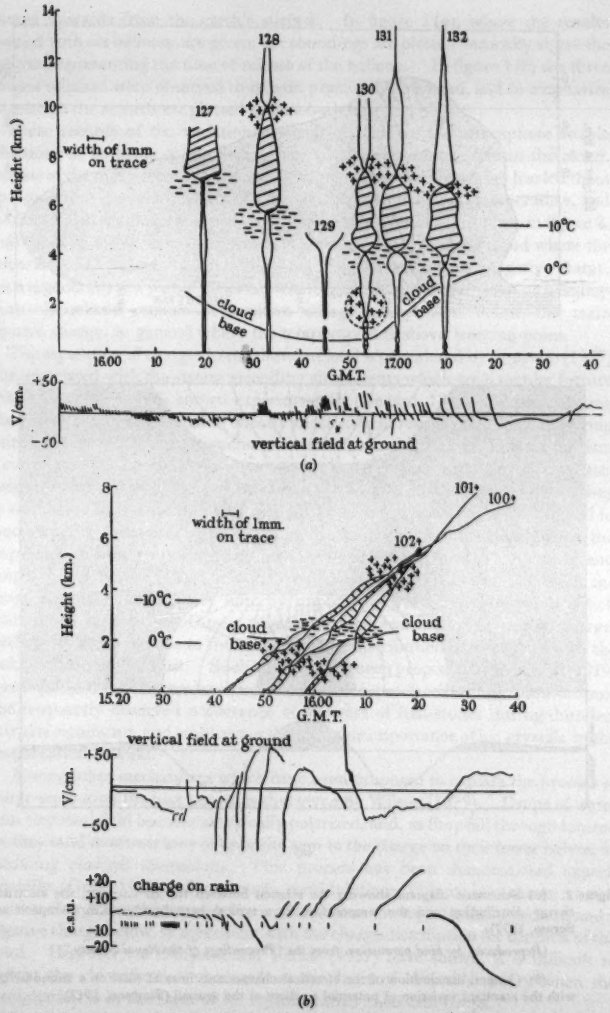
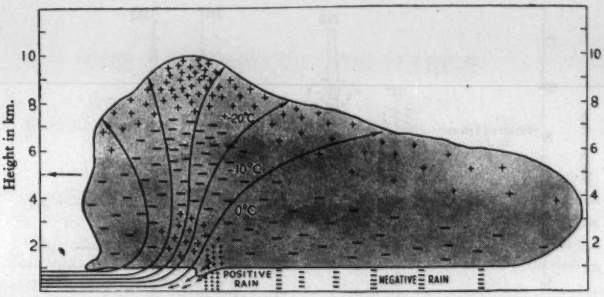
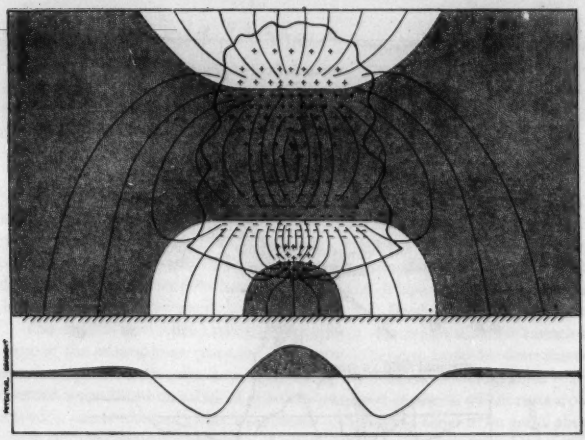


Figure 1. Analyses of the records obtained with alti-electrographs, carried by balloons, to show the variation with height above ground of the voltage gradient and charge distribution in two separate thunderstorms (Simpson and Robinson, 1941).

(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)



(*a*)



(*b*)

Figure 2. (*a*) Schematic diagram showing the relation between the air currents, the electrical-charge distribution and the temperature in a typical heat-thunderstorm (Simpson and Scrase, 1937).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

(*b*) General distribution of the electrical charges and lines of force in a thundercloud with the resultant variation of potential gradient at the ground (Simpson, 1942).

(*Reproduced by courtesy of the Royal Meteorological Society.*)

directed upwards from the earth's surface. In figure 1 (*a*), where the results obtained with six balloons are given, the soundings are plotted vertically above the abscissae representing the time of release of the ballons. In figure 1 (*b*) the three balloons released were observed to remain practically overhead, and to emphasize this relation the records are plotted against both time and height.

These records of the variation of electrical field in the atmosphere enable deductions to be made of the distribution of electrical charge within the cloud. Analysis of the many records obtained by Simpson and his colleagues has led them to propose that the relation between the air currents, the cloud temperature, and the charge distribution in a typical thundercloud is of the form shown in figure 2. In all thunderclouds there is a positive charge near the top of the cloud where the temperature is less than $-10°$ C. Beneath this positive charge is a negative charge, which is generally in a part of the cloud where the temperature is below the freezing-point. Localized regions of positive charge are present below the main negative charge, in general where the temperature is above freezing-point.

The separation of charges in the thundercloud is considered by Simpson (1927) to be associated with the strong ascending air currents which are a regular feature of thunderclouds. It is known experimentally (Lenard, 1892) that the breaking of large water-drops into smaller ones by an air current results in the drop acquiring positive charge, while a negative charge is carried away by the air. This mechanism is considered to account for the electrical separation near the cloud base. Positively-charged rain-drops accumulate in the base of the cloud, where they are suspended by the ascending air currents, and the negative charge is attached to cloud particles which pass upwards. In the upper regions of the cloud, where the temperature is below freezing-point, precipitation is in the form of ice crystals, and Simpson with Scrase (1937) and with Robinson (1941) have suggested that the charge separation takes place here by impacts between the ice crystals which result in the crystals requiring a negative charge and the air a positive charge. Settling of the ice relative to the air results in a distribution of electricity with the positive charge uppermost. Such a process has been proposed by Simpson (1919) to account also for the atmospheric electrification produced during snow-storms. The frequently observed occurrence of showers of hail-stones during thunder-storms is significant, and emphasizes the probable importance of ice crystals in the electrification process.

Among other mechanisms which have been proposed to explain the process of charge separation the best known is that given by Wilson (1929). Drops of water in an electrical field become electrically polarized, and, as they fall through ionized air, they tend to attract ions of opposite sign to the charge on their lower halves, so becoming charged themselves. This process has been demonstrated experimentally (Gott, 1935). From such considerations Wilson has shown that a positive charge may be expected to accumulate at the top of a thundercloud and a negative charge below, in agreement with the charge distribution for the bulk of the cloud. However, as Simpson and Scrase (1937) have shown, it is difficult to explain the electrification observed on the basis of Wilson's theory when the precipitation is in the form of ice crystals rather than rain-drops.

In observations of the rainfall occurring during thunderstorms, it is found

(Simpson and Scrase, 1937) that on the average a net positive charge is brought down from the cloud. Heavy rain occurs when the active centre is overhead, and is usually positively charged, as is to be expected on the basis of Simpson's proposals. The more moderate rain falling during the later stages of the storm may be either positively or negatively charged.

Measurements of the voltage gradient at the ground during the passage of the thundercloud overhead also appear to confirm the general picture put forward by Simpson. In figure 3 is shown the percentage frequency with which a positive field is observed for a large number of storms (Simpson and Robinson, 1941). Positive fields predominate as the storm approaches and recedes. The field is generally negative when the thundercloud is overhead, but there is an increased frequency of positive fields near the centre of the storm, in agreement with what one would expect for the charge distribution proposed in figure 2(*b*).

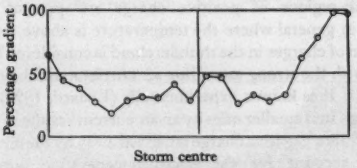


Figure 3. Frequency of occurrence of positive potential-gradients at the ground (Simpson and Robinson, 1941).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

The magnitude of the voltage gradients at the ground and in the atmosphere below and in the thundercloud is found to be no more than about 100 volts per cm. (Simpson and Robinson, 1941) as compared with a fine-weather field at the ground of the order of 1 volt per cm. The voltage at a cloud base of height 2 km. may then be considered to be approximately 20 million volts. A model of the charge distribution in the thundercloud which gives rough agreement with the observed voltage gradients has been proposed by Simpson and Robinson (1941). It consists of three spheres with their centres on a common vertical axis, and arranged as follows:—

(i) Lower positive charge: 4 coulombs in a sphere of 0·5 km. radius centred at a height of 1·5 km.

(ii) Main negative charge: 20 coulombs in a sphere of 1·0 km. radius, centred at a height of 3·0 km.

(iii) Upper positive charge: 24 coulombs in a sphere of 2·0 km. radius, centred at a height of 6·0 km.

§3. THE DEVELOPMENT OF THE LIGHTNING DISCHARGE

(i) *The lightning discharge from cloud to ground*

(a) *Photographic studies of lightning discharges.* Evidence of the composite nature of lightning discharges was obtained in 1890 by Hoffert, who photographed a discharge with a moving camera. He found that the discharge consisted of a

number of separate strokes which occurred at different intervals of time, of the order of ten milliseconds, and that the strokes followed the same track in space. This observation of the multiple strokes forming a single discharge led within a few years to further studies by Walter (1903) and Larsen (1905). In one photograph obtained by the latter a single lightning-discharge was resolved into forty separate strokes, the total duration of the discharge being about six-tenths of a second.

In order to determine the direction and speed of propagation of the strokes of a lightning discharge a special camera was devised by Sir Charles Boys (1926).

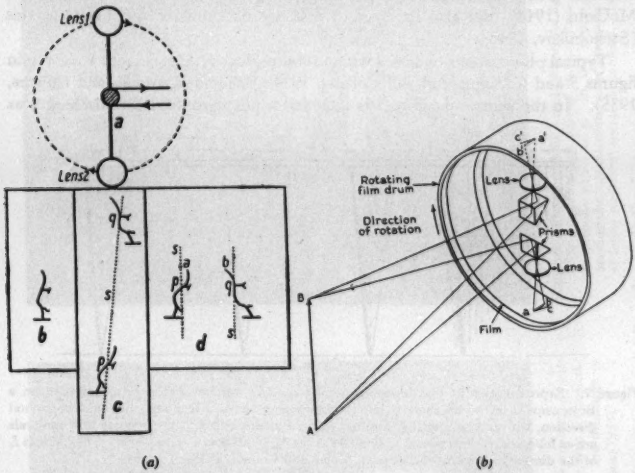


Figure 4. Schematic diagrams of Boys cameras:

(*a*) With moving lenses and stationary film (Schonland, 1938 b). (*Reproduced by courtesy of the University Press, Oxford.*)

(*b*) With moving film and stationary lenses (McEachron, 1939). (*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

This camera has two lenses, arranged to rotate at opposite ends of a diameter of a circle, as shown in figure 4 (*a*). Images of a lightning discharge are recorded on a stationary film behind the lenses. There is then a distortion of the photographed image resulting from the relative movement between lens and film, and a distinction can be made, therefore, between any portion of the discharge which becomes luminous before another portion. In a later form of the camera (Boys, 1929), shown diagrammatically in figure 4 (*b*), the lenses are fixed and the film is mounted on the inside of a circular rotating drum. A pin-point of light moving with uniform speed along a straight line from *A* to *B* would be recorded on the rotating film by the lines *ac* and *a'c'*, which are deflected from the lines *ab* and *a'b'* which would

have been obtained with a stationary film. The speed of development of the light from *A* to *B* can then be determined, provided the rotational speed of the film is known.

The first successful results with cameras of the above types were obtained by Schonland and his associates in South Africa (Schonland and Collens, 1934; Schonland, Malan and Collens, 1935, 38; Malan and Collens, 1937; Schonland, 1938 b; Schonland, Hodges and Collens, 1938). Subsequently similar records have been secured in the U.S.A. by Workman, Beams and Snoddy (1936), Allbright (1937), Holzer, Workman and Snoddy (1938), McEachron (1939) and McCann (1944), and also in the U.S.S.R. by Stekolnikov and his associates (Stekolnikov, 1940).

Typical photographs obtained with a rotating-lens type of camera are given in figures 5 and 6 (Schonland and Collens, 1934; Schonland, Malan and Collens, 1935). In the centre of figure 5 is included a photograph of the discharges as

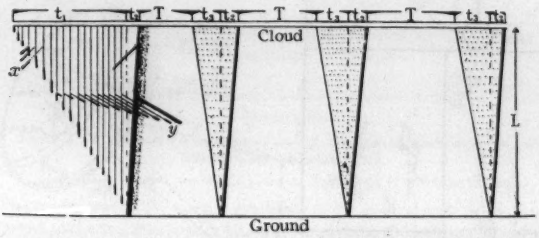


Figure 7. Representation of the development of a straight vertical discharge as recorded on a Boys camera, in which there is relative movement between lens and film in a horizontal direction, with time increasing from left to right of the record. The average time intervals are as follows : t_1, 0·01 sec. ; t_2, 0·00004 sec. ; t_3, 0·001 sec. ; T, 0·03 sec. The length L of the discharge is 2 km. (Schonland, Malan and Collens, 1935).

(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)

recorded by a stationary camera. In the centre of figure 6, which contains records of two separate discharges, denoted by the numbers 66 and 67, are shown moving photographs taken with another Boys' camera with a lower speed of relative movement between lens and film. The latter records enable the true sequence of the strokes to be determined more exactly than is sometimes possible with the higher-speed camera. The letters *a*, *b*, *c*.... denote the sequence of the strokes for the discharges in both figures 5 and 6.

From the analysis of the above and other similar photographs it has been possible to determine the general mechanism of lightning discharges (Schonland, 1938 a). This is shown schematically in the drawing of figure 7, which represents an idealized lightning-discharge developing along a vertical track between cloud and ground, as recorded on a film moving transversely to the direction of the discharge. The process is initiated by a streamer which develops downwards from the cloud in a series of steps which are separated by time intervals varying between

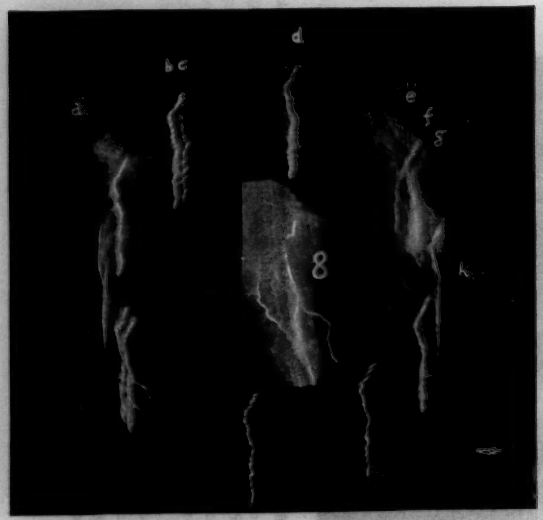


Figure 5. Photograph of a multiple-stroke lightning discharge, recorded with a rotating-lens type of Boys camera. Dart leaders are seen to precede strokes b and d. A stationary-camera photograph of the discharge is mounted in the centre of the picture (Schonland and Collens, 1934).

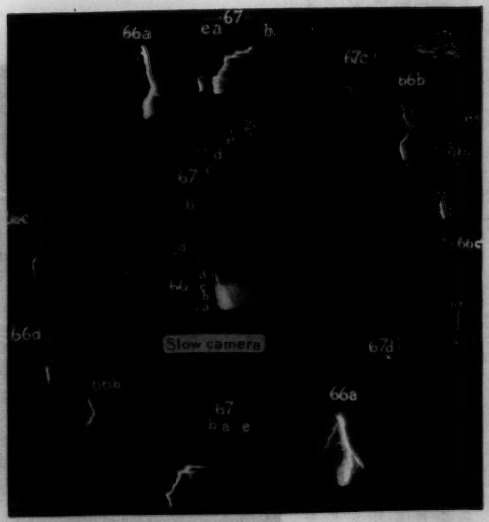


Figure 6. Photograph of a multiple-stroke lightning discharge, recorded with a rotating-lens type of Boys camera. A stepped leader is seen to precede stroke 66 a. In the centre of the picture is shown another photograph of the same discharge as recorded with a camera of lower rotational speed (Schonland, Malan and Collens, 1935).

(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)

To face page 330

Figure 8. Photograph obtained with a Boys camera of the development of a stepped leader-stroke (Schonland,[1] Malan and Collens, 1935).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

about 15 and 100 microseconds. Each step is revealed in the photograph by a sudden increased luminosity of the channel between the tip of the streamer and the cloud. The freshly ionized air at the tip of the streamer is more brightly luminous than the remainder of the channel. The speed with which the process, known as the stepped leader-stroke, approaches the ground is generally between 1×10^7 and 5×10^7 cm. per sec., with a usual value of about $1{\cdot}5 \times 10^7$ cm. per sec. The speed of the individual step-streamers is of the order of 5×10^9 cm. per sec.

In the photograph of figure 6 the development of a stepped leader-stroke is seen to precede the stroke 66 *a*. An enlargement of this photograph is given in figure 8.

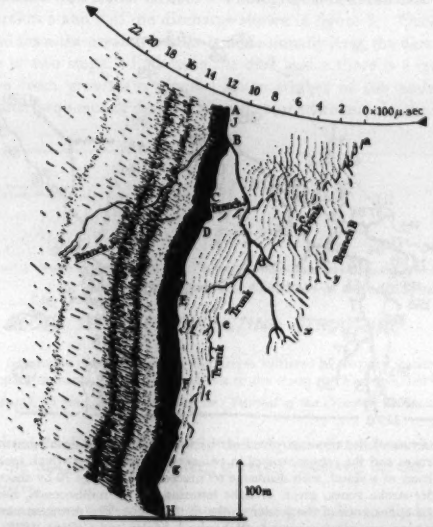


Figure 9. Diagrammatic analysis of a stepped leader-stroke, illustrating the development of the individual step-streamers (Schonland, Malan and Collens, 1935).

(*Reproduced, by kind permission, from the Proceedings of the Royal Society.*)

On account of the weak luminosity of the leader-stroke the details of its structure are not too clearly distinguishable. This is true of most photographs of stepped leader-strokes and, in some cases, no leader stroke has been recorded photographically, though confirmation of its presence is given by measurements of the electric-field variations which it produces at the ground.

An analysis of the fine structure of a stepped leader-stroke is shown in figure 9 (Schonland, Malan and Collens, 1935). Each step is seen to follow the path traced out by the preceding step, and branches which occur during the development of the leader-stroke advance by the same stepped process. Details of the time

taken for the leader-strokes and the return strokes to develop are given in figure 10 for two separate discharges. The times for the leader-stroke are in milliseconds, and those for the return stroke are in microseconds.

The above type of stepped leader-development is observed to precede the first stroke of a discharge in about 65 per cent of Schonland's records. In most of the remaining records the stepped leader-stroke is observed to develop more rapidly in its initial stages, with long bright steps (Schonland, Malan and Collens, 1938). Its effective velocity is greater than 6×10^7 cm. per sec., and in one case a velocity of $2{\cdot}6 \times 10^8$ cm. per sec. has been recorded. After the leader-stroke has proceeded part of the way between cloud and ground its speed suddenly falls, to a value which

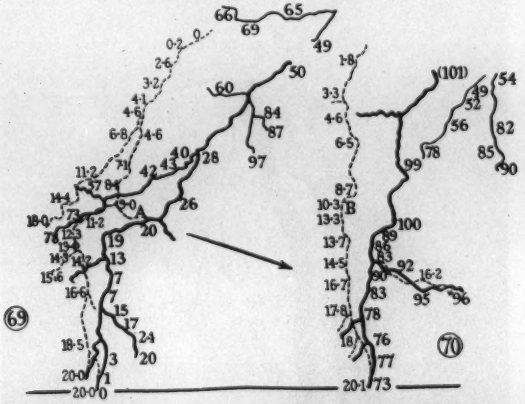


Figure 10. Analyses of the times involved in the various stages of development of the stepped leader-strokes and the return strokes in two separate discharges which took place from the same portion of a cloud, with discharge 69 preceding discharge 70 by about 0·0018 second. The leader-stroke times, given in light lettering, are in milliseconds, measured from the moment of appearance of the leader-stroke of flash 69. The return stroke times, given in heavy lettering, are in microseconds (Schonland, Malan and Collens, 1935).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

is usually about 1×10^7 cm. per sec. The step-length and the luminosity of the leader-stroke also decrease, and the process can only be photographed with difficulty. In some instances, during this second slow stage of the development, the leader-channel is brightly illuminated by occasional step-streamers, the time intervals between which may be as large as 0·01 second. Similar records, in which the discharge appears to proceed to ground in several large steps, have been obtained by Walter (1935) and by Workman, Beams and Snoddy (1936).

When the leader-stroke reaches ground, the main or return stroke commences to travel up the pre-ionized channel which has been established between the cloud and ground by the passage of the leader-stroke. The return stroke proceeds with a

speed of between about 2×10^9 and $1{\cdot}5 \times 10^{10}$ cm. per sec. The higher speeds are attained in its earliest stages near the ground. The photographed intensity of the return stroke is invariably much higher than that of the leader-stroke, and it is during the return stroke that large current flows are observed.

Subsequent to the passage of the stepped leader-stroke and the associated return stroke there may an interval of several hundredths of a second, after which a second leader and return stroke takes place, as shown in figure 7. This leader to the second stroke of the discharge generally consists of a streamer which develops in a single flight from cloud to ground, at an average speed of about 2×10^8 cm. per sec., and is termed a dart leader-stroke. Photographs of these dart leaders are recorded for strokes *b* and *d* of the discharge shown in figure 5. Occasionally, if the time interval since the previous stroke is exceptionally long, the dart leader may develop in one or two steps. Following the dart leader there is a return stroke which develops from ground to cloud. Later strokes of the same discharge consist of a similar sequence of dart leader and return stroke. In the majority of

TIME
INITIAL UPWARD STEPPED LEADER
SUBSEQUENT DISCHARGES WITH DOWNWARD CONTINUOUS LEADERS
HIGH SPEED BOYS CAMERA PHOTOGRAPH OF THE SAME STROKE
STILL CAMERA PHOTOGRAPH OF STROKE
CLOUD TO TALL CONDUCTING STRUCTURE

Figure 11. Schematic diagram showing discharges initiated by stepped leader-strokes developing upwards from high structures to the cloud (McEachron, 1939).
(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

cases each of these strokes follows closely the original channel developed by the first stroke of the discharge. Branching is rarely observed on the second and subsequent strokes.

The sequence of events described above refers generally to discharges taking place in open country. A pronounced difference is observed for lightning striking high buildings, as shown by the experiments of McEachron (1939) and later of McCann (1944) in the U.S.A. The discharge is found to be initiated, in most cases, by a stepped leader-stroke which develops upwards from the building to the cloud, as indicated in figure 11. The speed of the upward-developing leader-stroke, is about the same as that for the downward-developing leader-stroke, and the time intervals between the steps are of the same order of magnitude. But, unlike the discharge to open country, the initial stepped leader is not followed by a return stroke. After the leader-stroke has entered the cloud the luminosity of the channel persists, in what is known as a continuing stroke, during which current continues to flow. Downward-developing dart leaders are invariably observed preceding subsequent strokes of the discharge.

The majority of the photographs of lightning discharges have been obtained at night, as the usual photographic method requires that the shutter of the rotating camera should be open continuously before the flash occurs. Schonland (1938 c) has devised a method whereby the camera shutter is opened by a mechanism triggered by the electromagnetic pulses which are radiated during the stepped leader process. This enables the shutter to open in sufficient time to record the later development of the stepped leader-stroke as it approaches the ground, and also the return stroke and subsequent strokes. The shutter is arranged to close automatically after 0·5-second exposure. An alternative method, which gives less satisfactory results, makes use of a chlorine-bromine filter in conjunction with a quartz-fluorite lens. With this arrangement little fogging of the film takes place with the shutter open continuously in daylight, whereas light of short wave-lengths from the lightning discharge is transmitted by the filter-lens system, and photographs can be obtained in this manner of discharges at distances up to about three miles. Photoelectric recording of the light emitted by lightning discharges, at distances as great as twenty miles, has been successfully employed by Schonland (1938 c).

(b) *The mechanism of leader-strokes.* The luminosity of the various processes comprising the lightning discharge is greatest near the tips of the streamers or leader-strokes. This concentration of light may be considered to be caused by the greater fields existing at the tips, as compared with the stems, of the streamers, so that greater numbers of atoms and molecules are raised to excited states by electron impact. In general it is only the tips of the leader-strokes which are sufficiently bright to be photographed. In the dart leader the emitted light is largely confined to a distance of about 50 m. behind the tip and is only weakly discernible at greater distances behind the tip. Similar considerations apply to the step streamer. In the return stroke the emitting region is longer, but the luminosity near the ground has fallen appreciably before the bright tip of the stroke has reached half-way to the cloud.

Schonland (1938 a) has postulated that the stepped leader process is preceded by a faintly luminous " pilot " streamer, carrying a current of the order of several amperes and which travels continuously with a speed equal to the effective rate of advance of the leader-stroke. The weakly ionized channel developed by the pilot streamer becomes unstable as the streamer advances, and after a period there is a rapid readjustment of the ionization in the channel, by which it is rendered more brightly luminous and more highly conducting. This process represents the passage of the step streamer, which advances rapidly down the channel to catch up with the advancing pilot streamer.

The existence of the pilot streamer still requires experimental confirmation. However, it is reasonable to suppose that even if it does exist, and there is circumstantial evidence in its favour, its luminosity would be so weak as to render it extremely difficult to photograph. By consideration of its existence, Schonland (1938 a) has been able to put forward an explanation of the remarkable regularity which is observed in the time intervals between the successive step-streamers, which, in 90 per cent of the stepped leaders observed, lie between 50 and 90 microseconds. The explanation is based on consideration of the decrease of

electron density in the leader channel caused by electron capture and recombination processes. A decrease in electron density necessitates an increase in field strength if the current is to be maintained in the channel, and ultimately the field attains a value sufficient to cause breakdown and, therefore, fresh ionization in the channel, so that a step streamer is formed. This explanation has been extended quantitatively by Loeb and Meek (1940), who have estimated the probable decrease of ionization in the channel between steps in terms of recombination processes. An alternative proposal has been put forward by Bruce (1941, 44), who has considered the corona losses from the leader channel. As the channel lengthens an increased current is required to supply the increased corona-loss, and Bruce has suggested that, when this current reaches a particular value, a sudden transition from a glow to arc conditions will take place. This suggestion has been criticized by Meek (1941) on the grounds that the experimental data on which it is based refer to observations of the glow-to-arc transitions across short gaps between solid electrodes, for which it is known that electrode effects are important. Whereas, in the case of the lightning leader-stroke, where no solid electrodes are present, there is some doubt as to the possibility of a sudden transition from a glow to an arc for a particular current in the channel.

The usual speed of advance of the stepped leader-stroke, which is about $1{\cdot}5\times10^7$ cm. per sec., is closely that of the observed velocity of electrons in the minimum ionizing field in air. The much higher speed of about 5×10^9 cm. per sec. which is attained by a step streamer may be explained by the presence ahead of its tip of a pre-ionized path in which appreciable ionization remains from the passage of the preceding streamer. Schonland (1938 a) has given an expression for the velocity V of advance of a streamer along a pre-ionized path as

$$V = N^{1/3}\bar{v}\,d, \qquad \ldots\ldots(1)$$

where N is the pre-existing electron density in electrons per cm^3 in the ionized path ahead of the tip, $\bar{v}$ cm. per sec. is the electron drift-velocity in the field immediately ahead of the tip, and d cm. is the distance ahead of the tip in which the field exceeds the critical value for ionization by electron impact. Clearly values of V very much greater than the electron drift-velocity $\bar{v}$ may be attained in the presence of high values of N. Similar considerations apply in the case of the return stroke and the dart leader, but, because of the greater age of the ionized channel preceding the development of the dart leader, the speed of the latter is very much less than that of the return stroke or the individual steps of the initial stepped leader-stroke.

The final explanation of the mechanism of the leader-stroke has not yet been given. No satisfactory theory has yet been put forward from which calculations can be made, in terms of fundamental physical processes, of the voltage gradients required to cause the propagation of leader-strokes. Before a complete explanation can be given it would appear that further experiments are necessary in order to determine more accurately the processes involved in the formation of a leader-stroke and the conditions of ionization, of current and of voltage distribution in the leader-channel. Because of the random and uncontrolled nature of lightning discharges the progress of experimental work is necessarily slow, but much relevant and

useful information may be gained from studies in the laboratory on the development of long sparks. This was early realized by Schonland, who, after his observations of leader-strokes preceding lightning discharges, made a photographic study with a rotating-film camera, in conjunction with Allibone (Allibone and Schonland, 1934), of sparks across a 100-cm. gap. The results showed the presence of a leader-stroke followed by a main stroke. The studies on spark discharges were continued by Allibone and Meek (1938 a, 38 b), who found that many of the characteristics of lightning discharges could be obtained in the laboratory. The stepped process was observed for both positive and negative leader-strokes, but with certain differences as shown in the photographs given in figures 12 and 13 for sparks between a high-voltage point and an earthed plane. For the discharge from a positive point, shown in figure 12, the leader-stroke starts from the point and develops across the gap to the earthed plane, when it is followed by the return stroke, which retraces in detail the channel traced by the leader-stroke. Branches traced by the leader-stroke are retraced by the return stroke. The discharge between a negative point and an earthed plane is initiated by a leader-stroke developing in a stepped manner, as shown in figure 13. On the occurrence of the final step a positive leader-stroke develops from the earthed plane to meet the descending negative leader in the mid-gap region. The velocities of the leader-strokes observed in the laboratory vary from about 10^8 cm. per sec. upwards, and depend to some extent on the external circuit conditions. In the case of the positive leader-stroke the tip of the leader is generally observed to develop continuously, reinforced by step streamers, but there does not appear to be a continuous streamer joining the tips of the successive negative streamers in the negative leader-stroke. The existence of the pilot streamer, as postulated by Schonland (1938 a) for the lightning leader-stroke, is not confirmed therefore by the present laboratory evidence on spark development from a negative electrode, and it is possible that the final explanation of the leader-stroke may not require the pilot-streamer hypothesis.

The upward development of the positive leader-stroke from the earthed plane for the negative discharge is of interest as this is not observed in the case of leader strokes preceding lightning discharges in open country, where, in the majority of cases, the lightning discharge conveys a negative charge to ground. Considerations of the electric-field distribution in a point-plane gap show that the field at the plane is too small to produce ionization at the plane. However, in an explanation of the upward development of the positive leader from an earthed plane, Meek (1942) has suggested that it is produced as a result of the high field induced at the plane by the negative charge sprayed into the gap by the descending negative leader. This view is generally confirmed by experimental evidence on sparks, where it is possible to obtain a measure of the quantity of charge involved in the leader process. As the approximate induced field at an earthed plane is inversely proportional to the square of the height of a charge above the plane, the field decreases rapidly with increasing height of a given charge of given magnitude. It is understandable therefore that upward-developing positive leader-strokes are less prominent for longer gaps, though they may develop when the descending negative leader approaches near to the plane. It is possible then that leader-strokes developing upwards from the ground may be present in lightning discharges, but that they are too short to be recorded under normal conditions where the camera is at a consider-

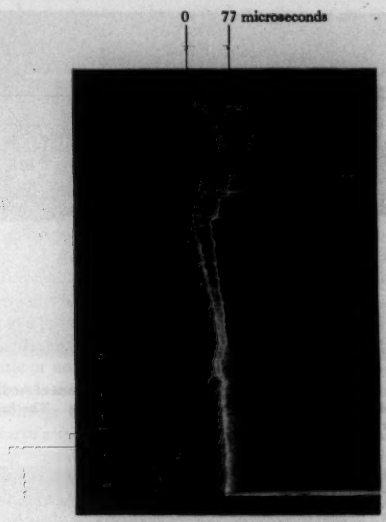


Figure 12. Photograph obtained with a moving-film camera of the development of a spark across a 130-cm. gap between a positive point and an earthed plane.

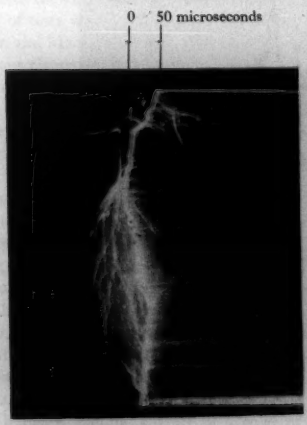


Figure 13. Photograph obtained with a moving-film camera of the development of a spark across a 100-cm. gap between a negative point and an earthed plane.

Figure 14. Photograph obtained with a moving-film camera of the development of a spark across a 100-cm. gap between a negative sphere, of 100 cm. diameter, and an earthed point projecting 50 cm. above the ground plane.

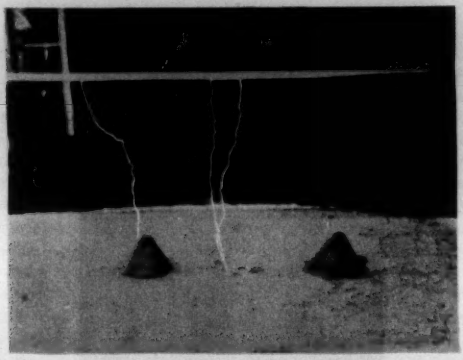


Figure 15. Sparks from a metal rod, of 1·6 cm. diameter, to a plane formed of sand, with two hills. A metal sphere of 12·5 cm. diameter is buried between the hills. The height of the rod above the plane is 32 cm.

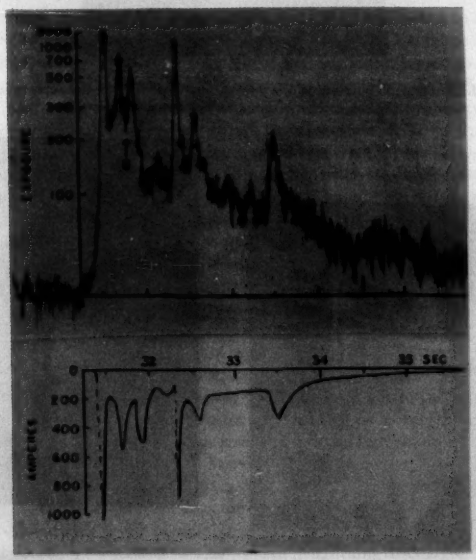


Figure 18. Comparison between the variation of intensity of the image of a discharge, as photographed on a moving-film camera, with an oscillogram of the current flowing in the discharge (Flowers, 1944).

(*Reproduced, by kind permission, from the* '*General Electric Review*'.)

able distance from the discharge. One still photograph, given by McEachron (1939), shows the presence of several streamers from the ground, of lengths about 200 cm., which may represent upward-developing leader-strokes.

With an electrode arrangement in which the high-voltage electrode is a large sphere and the earthed electrode a long point, upward-developing positive leader-strokes are observed in the laboratory to travel the whole distance across the gap from the point to the sphere, as shown in figure 14 (Allibone and Meek, 1938). The leader-stroke which is stepped may be likened to the upward-developing leader-strokes observed by McEachron (1939) for lightning discharges to high buildings.

Other studies related to the development of leader-strokes have been carried out by Loeb and his associates (Loeb and Kip, 1939), who have used oscillographic methods to observe the onset of corona from points and have obtained valuable information concerning the initiation and propagation of streamers.

The branching of a lightning discharge in open country appears to be affected by the nature of the terrain. In flat, bare country McCann (1944) reports that discharges show little or no branching, whereas in hilly, wooded country profuse branching is usually observed. It is probable that in the latter case there is an increased point-discharge current, which provides a positive-ion space charge blanket near the earth's surface, and encourages the development of branches from the descending negative leader-stroke. Similar results have been observed in the laboratory for a spark between a high-voltage negative-point and an earthed plate covered with a number of projecting points (Schonland and Allibone, 1931).

The development of the leader-stroke is clearly an important factor in governing the selection of the earthed object which is struck by the lightning discharge. High buildings, which may initiate leader-strokes, are naturally more prone to be struck than the surrounding country when thunderstorms are in the vicinity. A number of tests with models in the laboratory have been carried out by Wagner, McCann and MacLane (1941), who have studied the propagation of leader-strokes for various electrode arrangements which simulate on a small scale the conditions obtaining for lightning discharges. Consideration of the height of lightning rods and the degree of protection afforded to neighbouring objects have been studied, together with the influence of soil resistivity and ground contours. Other recent investigations of the protective value of lightning conductors, by means of scale models, have been carried out by Zalesski (1935), Akopian (1937), Matthias (1937), Matthias and Burkardtsmaier (1939), Wagner, McCann and Lear (1942). Golde (1941) has discussed the limited applicability of model tests in relation to the usual conditions obtaining during thunderstorms.

The localization of lightning discharges has been discussed by Goodlet (1937). It appears that the influence of soil conductivity is appreciable and that regions where there is a discontinuity in the geological formation are prone to be struck (Fritsch, 1941). A series of relevant experiments on sparks in the laboratory have been carried out by Stekolnikov and his associates (Stekolnikov and Yavorsky, 1936; Stekolnikov and Beliakov, 1936; Beliakov and Hanov, 1936; Stekolnikov, 1937 b). In one of these investigations (Stekolnikov, 1937 b) a metal ball was buried beneath the flat surface of a layer of soil, and two small hills of soil were

raised near the buried ball. The soil was made to form one electrode of a spark gap, and it was found that, on the application of impulses voltage to the gap, the sparks took place across the longer gap to the buried metal ball when the soil conductivity was low, in preference to the shorter gap to the hill-tops. When the soil conductivity was increased the sparks took place to the hills. A photograph obtained by Miranda (1937) for a similar arrangment is shown in figure 15, where sparks are seen to occur to a buried metal ball.

(c) *The return stroke.* The high speed of the return stroke is accounted for by the existence of the pre-ionized channel ahead of its advancing tip, as explained in

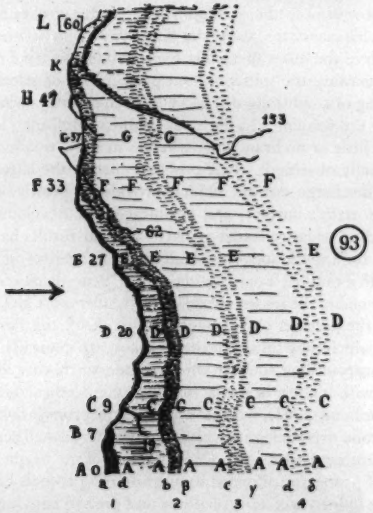


Figure 16. Diagrammatic analysis of the photographed structure of a return stroke. The figures at the leading edge denote the times, in microseconds, at which the return stroke first reaches adjacent points (Malan and Collens, 1937).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

the previous section for step streamers and dart leaders. A gradual decrease in speed takes place as the return stroke proceeds up the channel, probably because of the greater age of the channel nearer the cloud.

Changes in luminosity of the main channel have been studied by Malan and Collens (1937), and in figure 16 a photometric analysis of one of their photographs is given. They conclude that the main changes in luminosity, which may be associated with the current in the main channel, takes place when the tip of the return stroke reaches a branch. The luminosity of the channel between the ground and the branch then increases, while in its further progress up the channel the

return stroke proceeds with diminished luminosity. The branch is illuminated to an extent comparable with the portion of the main channel between the branch and ground. In an explanation of the change of luminosity at branching points Schonland has drawn attention to the relative ages of the ionized trails in the main channel and in the branches. This may be understood by reference to figure 17, which represents the time sequence of a vertical discharge with a single branch (Schonland, 1938 a). The light lettering denotes the time, in microseconds, at which the various parts of the main channel and the branch were formed. The heavy lettering denotes the age of the ionization at each point when the returns streamer reaches it. For simplicity the speed of the return stroke has been assumed

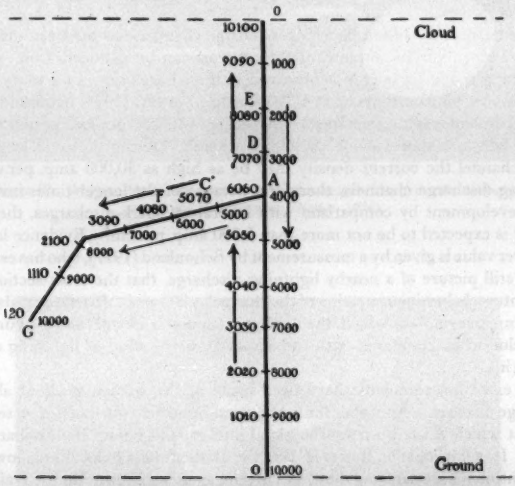


Figure 17. Analysis of the age of the ionized channel, for a hypothetical discharge in a vertical direction, with one branch (Schonland, 1938 a).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

to be uniform and to be 10^3 times that of the leader-stroke. At the end of the branch the age of the trail is little different from that at the base of the channel. The age at E on the main channel is 8,080 microseconds, whereas the age at F, which is equidistant from the branching point A, is only 4,080 microseconds. Therefore, when the return stroke reaches A the portion travelling along the branch develops faster than that along the main channel, and the branch is illuminated more brightly than the portion of the main channel above.

By consideration of the rotating-camera photographs in conjunction with synchronized current oscillograms of lightning discharges to the Empire State Building, Flowers (1943, 44) has been able to obtain approximate agreement

between luminosity and current. An example of the results is given in figure 18,* where a comparison is made between the current oscillogram and the luminosity as determined with a micro-densitometer from the rotating photograph of the discharge. While the general agreement between the two records is evident, there are difficulties in establishing the calibration scale for the estimation of currents from the luminosity records, as confirmed in similar measurements by McCann (1942a 44). Some improvement may be obtained by the use of a photoelectric electron-multiplier coupled to an oscillograph, which records directly the variation with time of the luminosity of lightning discharges, without the intermediate step of a photographic record; this arrangement has been employed by Meek and Craggs (1943) in the study of the light emission from spark discharges.

The current density in lightning-discharge channels has not been determined directly, but a certain amount of information can be deduced from evidence concerning spark discharges as observed in the laboratory. In a study of spark channels, carrying currents up to 100,000 amp., Flowers (1943) has found that the channel expands rapidly, at a speed of the order of 10^5 cm. per sec., until the current density falls to a value of about 1,000 amp. per cm^2 During the initial stages of the spark channel the current density may be as high as 30,000 amp. per cm^2 In lightning-discharge channels, therefore, because of the longer times involved in their development by comparison with laboratory spark-discharges, the current density is expected to be not more than 1,000 amp. per cm^2 Evidence in favour of a lower value is given by a measurement by Schonland (1937), who has estimated, from a still picture of a nearby lightning discharge, that the cross-sectional area of the intensely luminous portion of the channel is 200 cm^2 If a current density of 1,000 amp. per cm^2 is assumed, the discharge current is 200,000 amp., an unusually high value when compared with the observed distribution of lightning currents (see § 4 (ii)).

No exact measurements have been made of the voltage gradient along the discharge channel. An upper limit of several hundred volts per cm. is set by the gradient which exists between the cloud and ground before the discharge takes place. It is improbable, however, that the gradient falls to a value as low as that for a continuous arc carrying the same current, as an insufficient time is available for equilibrium to be established between the channel and the surrounding atmosphere. A guide to the probable magnitude is given in measurements of the voltage gradients along spark channels by McCann and Clark (1943), who find values about ten times those of arcs for the same current, at times up to 20 microseconds after the establishment of the spark.

(d) *Multiple strokes.* The reason for the multiple-stroke sequence observed in lightning discharges has been ascribed to the extension of the channel to new charge centres within the cloud. Schonland (1938a) has suggested that this extension may take place by the development of leader-strokes from the new charge centres towards the existing charge centre. Alternatively, as pointed out by Bruce and Golde (1942), the leader-strokes may develop from the channel to the new charge centres.

Another possible cause of multiple strokes is the repeated charging of the

* Facing page 327.

Figure 19. Moving-film photograph showing a change in the path of a discharge containing three strokes (Walter, 1936).

Figure 20. Still photograph of lightning discharge to ground and within the cloud.

To face page 330

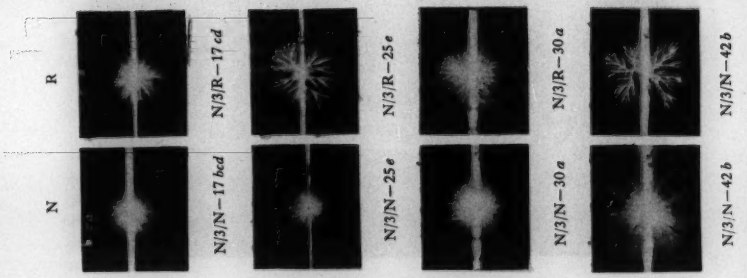


Figure 24. Klydonograms of voltages on transmission-line conductors, due to direct lightning strokes to the transmission line.

N = normal film, i.e. point electrode at high voltage-plane electrode earthed.
R = reverse film, i.e. plane electrode at high voltage-point electrode earthed.

(Perry, Webster and Baguley, 1942.)

(Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.)

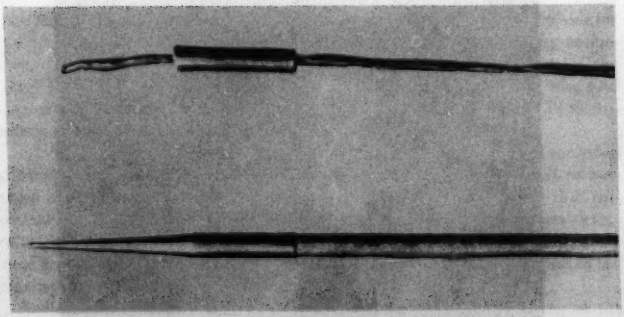


Figure 25. Hollow-copper lightning rods, initially identical in size: one unused, the other crushed by a lightning discharge (Humphreys, 1940).

(Reproduced by courtesy of the McGraw-Hill Publishing Company, New York.)

original cloud centre from which the discharge originally developed (Perry, 1944; McCann, 1944). Some evidence in favour of this process is given by McCann, who has drawn attention to two rotating camera photographs of multiple strokes, in one of which the time intervals between a series of eight strokes is remarkably regular, between 0·0093 and 0·0112 seconds. McCann suggests, as a result of laboratory data on multiple-stroke discharges in the laboratory, that the continuous charging of the one cloud centre is the responsible factor in such instances. Further, the rate of charging, which is considered to proceed at a fairly uniform rate, is in accord with the measurements of voltage gradient at the ground.

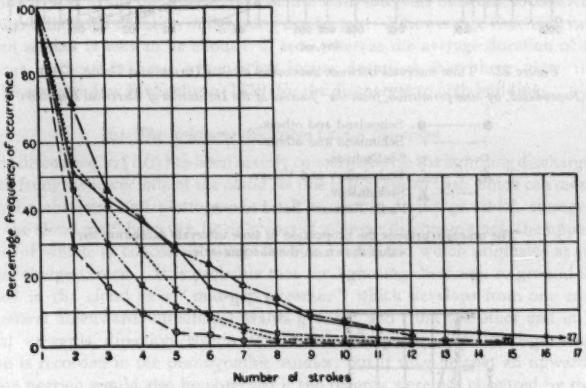


Figure 21. Number of successive strokes in lightning discharges (Bruce and Golde, 1942).
(*Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.*)

●——● Schonland and others.
O– – –O McEachron.
△–·–·–△ Stekolnikov and Valeev.
□–·–·–□ Noto.
×·····× Radio Research Board records.

The points represent the proportion of flashes in which the number of component strokes is equal to, or greater than, the value shown on the abscissa-axis.

While the component strokes of most lightning discharges take place along the same channel, exceptions are sometimes observed. A photograph obtained by Walter (1936) of a discharge containing three strokes is shown in figure 19. The first stroke on the left of the photograph is followed after 23 milliseconds by the second stroke, which retraces part of the path taken by the first stroke, but then develops along a separate track, the initial part of which has been formed as a branch of the first stroke. After a further 240 milliseconds the third stroke develops, and follows the path of the second stroke. Similar observations are reported by Schonland, Malan and Collens (1938), who consider that changes in the lower portions of the channels used by multiple strokes are to be associated with

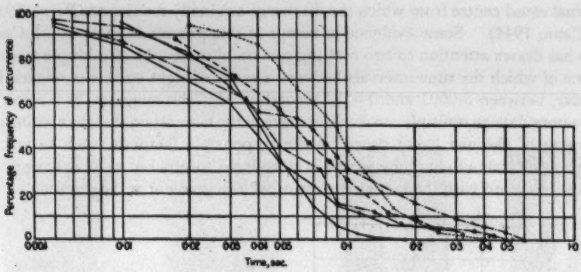


Figure 22. Time intervals between successive strokes (Bruce and Golde, 1942).

(Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.)

● ——— ● Schonland and others.
\+ ——— + Schonland and others.
○ – – – ○ McEachron.
△ – · – · – △ Stekolnikov and Valeev.
— · — · — ▲ Stekolnikov.
× · · · · · · × Radio Research Board records.

The points represent the proportion of time intervals exceeding the value shown on the abscissa-axis.

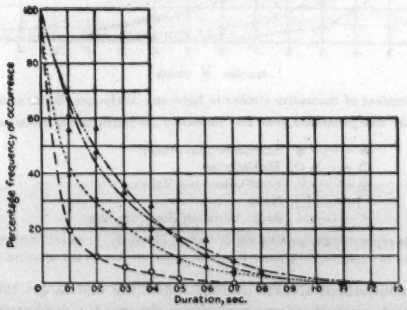


Figure 23. Total duration of lightning discharges (Bruce and Golde, 1942).

(Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.)

⊙ — — — ⊙ McEachron.
△ – · – · – △ Stekolnikov and Valeev.
▲ ——— ▲ Stekolnikov.
× · · · · · · × Radio Research Board records.

The points represent the proportion of flashes in which the duration shown on the abscissa-axis is exceeded.

long time intervals between the strokes, whereby the ionization existing in the channel has largely disappeared. A change in the upper portion of the channel may be attributed to the presence of a second charge-centre within the cloud.

A comparison of the number of multiple strokes in a discharge, as analysed by Bruce and Golde (1942) from the results of a number of workers in different parts of the world, is given in figure 21. In exceptional records, not given in this diagram, as many as 40 strokes were observed by Larsen (1905), and 42 strokes by Matthias (1929).

Further analyses given by Bruce and Golde (1942) are given in figures 22 and 23, and show respectively the distribution of the time intervals between successive strokes and the total duration of lightning discharges. The average time interval between strokes is seen to be about 0·07 sec., whereas the average duration of a discharge is about 0·25 sec. Somewhat longer durations than those given in figure 23 are found by McEachron (1939) for the discharges to high buildings.

(ii) *The lightning discharge within the cloud*

The discussion in § 3 (i) has been mainly concerned with the lightning discharge to earth from the lower side of the cloud, as this is the type of flash which can most readily be observed and photographed. The visible discharge which emerges below the cloud may, however, be only part of a longer discharge track, the upper portion of which is concealed within the cloud itself and which originates at an internal charge-centre. It is probable that the lightning discharge to ground is initiated in the cloud as a " mid-gap streamer " which develops from one end in a general downwards direction towards ground, and from the other end in a general upwards direction within the cloud. Only the downward-growing portion is recorded in the photographic studies, but it may be that an upward-growing portion would also be observed if the records were not obscured by the cloud opacity. In the case of multiple-stroke flashes it has been suggested (see § 3 (i)(*d*)) that nearby charge-centres may be involved in the strokes subsequent to the first so that the concealed portion of the discharge may vary both in direction and length during the progress of the discharge.

A large proportion of lightning discharges occur within the cloud itself, or may progress towards the earth without reaching it. Discharges of this latter type have been termed " air discharges " (Schonland, Malan and Collens, 1935), and appear to be caused by discharges to heavy space-charge concentrations below the cloud (Schonland, 1938 a). Alternatively, the development of the leader-stroke may be arrested by the collapse of the electric field, possibly because of the occurrence of other neighbouring discharges. The presence of discharges wholly within the cloud, and which, therefore, cannot directly be seen, is known by observation of the associated phenomena, such as field changes, general illumination of the cloud, and thunder. In certain types of discharge, however, part, at least, of the track is directly visible. Shipley (1940) has attempted to classify nine different types of discharges frequently seen, of which four consist of discharges to ground and five consist of discharges within or about the cloud. Two of these latter types usually takes place high up in the cumulus cloud and appear to involve localized charges of a portion of the cloud only. The other three types are shown

as much longer discharges whose main direction is horizontal and which appear to traverse the major part of the cloud structure. Where these discharges occur near the lower edge of the cloud, they may be branched or non-branched.

The relative proportions of cloud discharges and ground discharges are not accurately known but they appear to vary widely in different localities. Schonland (1928 a) for storms observed at Somerset East, South Africa, found that the ratio of discharges in the cloud to discharges to ground varied from 50/1 to 5/1, with an average value of 10/1. Simpson (1909), when measuring the charge on rain from thunderstorms at Simla, observed that "most of the lightning discharges" took place within the cloud between the base and the summit. Halliday (1932) measured the field changes due to lightning discharges in Johannesburg and observed the type of discharge in 560 out of a total of 3,015 field-change records. Out of the *observed* discharges 277 were in the cloud and 283 were to ground, but an analysis of the total number of field changes would indicate that the ratio of total cloud discharges to total ground discharges was of the order of 1·7/1. The discrepancy is probably due to the greater difficulty of observing cloud discharges. According to Golde (1945), Whipple and Scrase (1936) observed that the ratio of cloud discharges to ground discharges at Kew was about 0·7/1.

The large differences in the observations by different workers of the ratio of cloud discharges to ground discharges are not fully understood. It is probable that they are influenced by the average height of the cloud base above ground, as, with decreasing cloud height, the number of discharges to ground may be expected to increase. There appears to be some correlation (McCann, 1944) in the results obtained by various workers between the cloud height and the relative numbers of cloud and ground discharges. However, other factors may be important, and it is anticipated that differences are to be expected in the results for thunderstorms in tropical and temperate zones, where the thermodynamical conditions producing the thundercloud vary, and where the distribution of charges in the cloud and the distance between charge centres may differ.

McCann (1942 b) describes a number of cloud-to-cloud discharges photographed by a moving-film camera of sufficient resolution to separate the individual strokes. Some of his records show vertical or inclined discharges clearly involving the upper positive and the lower negative charge-centres of the cloud. The discharges are multiple in character and generally have a larger number of components, shorter time-intervals between components, and a shorter total duration of the flash than is the case for discharges to ground (McCann, 1944). One photograph shows a multiple-stroke air-discharge with pronounced branching (McCann, 1942 b), and a second shows a multiple-stroke discharge originating at the top of the cloud and penetrating the cloud to reach the earth (McCann, 1942 b; McCann and Morgan, 1943). Evans (1943) shows a still photograph of a ground discharge which was seen by visual observation to be linked to the discharge in the cloud-top above it. The photograph given in figure 20 shows both a cloud discharge and a ground discharge, which may be associated, but there is no evidence concerning the sequence of the processes involved.

The photographic analysis of cloud discharges is not in all cases sufficiently exact to show the mechanism of their propagation, but this may be deduced from

consideration of the electrical-field changes produced. For "air-discharges" Schonland, Malan and Collens (1935) showed that the discharge streamers progress in a series of steps in a similar manner to that of the first stroke of a multiple-stroke discharge, but that no return strokes occur. In measuring the field changes due to discharges within the cloud, Schonland, Hodges and Collens (1938) found that the usual field-change from such a discharge was a simple slow rise to a final maximum, the rise carrying superimposed pulsations of the same kind as those found for the stepped leader process and with the same time-interval between pulsations. There was an entire absence of the sudden "b" change of field which occurs with a stroke to ground at the time of the development of the main return stroke (see § 5). From this observation, Schonland with Hodges and Collens (1938) and with Malan and Collens (1938) deduced that for discharges within the cloud the mechanism was the same as for air discharges, i.e. a stepped leader process but no return stroke.

A further process is demonstrated by the measurement of the wave form of atmospherics by Lutkin (1939) and Schonland, Elder, Hodges, Phillips and van Wyk (1940). In addition to the group of wave forms associated with discharges to ground (see § 5), two other groups (II and III) have been observed to be due to discharges within the cloud. Group II is associated with the stepped leader process without a return stroke. Group III includes wave forms similar to those obtained from discharges to ground but of smaller amplitude, and separated by short quiet intervals. Schonland and his colleagues suggest that these arise from multiple discharges initiated by dart leaders.

The total field-changes due to discharges within the cloud have been found by Wormell (1939) to be of the same order of magnitude as those for discharges to earth so that the range of charges to be neutralized is the same in the two cases.

(iii) *Bead and ball lightning*

Bead lightning, which is observed fairly frequently, is caused by the persistence of luminosity of the discharge channel, which remains visible for longer times in certain sections than in others. Consequently the channel, before it disappears, may appear to break up into a number of detached light-sources. This subject has been discussed in a recent summary by Wolf (1943). In a description of the phenomenon, an observer (Beadle, 1936) has estimated the presence of 20 to 30 luminous beads remaining in the path of the discharge and spaced at about 2 feet. The diameter of these beads of light was estimated at about 3 inches, and their life-time as about 1/2 sec., as compared to a duration of about 1 sec. for the preceding main discharge. The persistence of these local regions of light is witnessed by the streaks observed in rotating camera photographs, as, for instance, in figure 19. It frequently occurs at bends in the channel, and has been explained by various authorities as an optical effect caused by the end-on view of the channel in the region of the bend, so that an effectively deeper length of ionization is in the line of sight. The persistence of the light appears to be too long to be accounted for in this manner, and it is probable that it is caused by more fundamental, as yet unexplained, processes (Craggs and Meek, 1946).

The evidence concerning ball lightning, sometimes termed a thunderbolt or

fireball, is less definite, and many persons are sceptical as to its existence. But there is an appreciable amount of information available from the various accounts which have been given by the various observers, and the number of these independent accounts, together with their general agreement, would appear to substantiate ball lightning as a true phenomenon. Brand (1923) has collected some 600 accounts of ball lightning, from about one-third of which he has been able to deduce the principal characteristics of the phenomenon. Other accounts have been investigated by Humphreys (1936), who concludes that in many of the observations the phenomenon can be explained as a subjective effect, influenced by persistence of vision. The subject has been discussed more recently by Goodlet (1937). In general ball lightning appears as a reddish luminous ball of 10 to 20 cm. diameter, which lasts for several seconds, and which sometimes, though not always, appears after an ordinary lightning discharge. They may disappear silently or with an explosion. Photographs of unusual discharges which may represent ball lightning have been given by Holzer and Workman (1939).

Various theories have been put forward to account for ball lightning. Thornton (1911) has suggested that it is formed largely of a mass of ozone and dissociated oxygen, produced by a normal lightning discharge, and that this reverts suddenly to molecular oxygen with explosive violence. The energy balance is calculated to be such as to account for the process. In a more recent theory, proposed by Neugebauer (1937), it is shown that the stability of a gaseous sphere containing electrons and positive ions can be explained in terms of quantum mechanics, by consideration of exchange forces, if the electron density is sufficiently high, as may be the case in lightning discharge channels. The phenomenon has also been discussed theoretically by Frenkel (1940).

§4. MEASUREMENTS OF LIGHTNING DISCHARGE CURRENTS

(i) *Methods of measurement*

As it is not possible to judge beforehand, except in certain particular cases, where a lightning discharge will take place, the problem of measuring the current in the discharge is a difficult one. Accordingly the early methods of measurement employed the simplest devices, as these could be installed, with a minimum of maintenance, at many points, for example along an electric power transmission line, in the hope that at least a few of the devices would yield results. For certain high structures, where the chances of a direct lightning stroke is comparatively high, it has been possible to use more elaborate instruments with a consequent improvement of accuracy and an extension of the data obtained.

In some of the earlier devices for the measurement of lightning currents use was made of the phenomenon known as the Lichtenberg figure, which is the pattern formed on a photographic film or plate by a surface corona discharge (Merrill and von Hippel, 1939). The usual form of this type of instrument is known as the klydonograph (Peters, 1924; Cox and Legg, 1925; Lee and Foust, 1927; Lewis and Foust, 1930; Berger, 1936; Cox, McAuley and Huggins, 1927; Candler, 1940), and may be used with a resistance shunt to measure a current surge or with a potential divider to measure a voltage surge. In each case the resultant surge-

voltage is applied between two electrodes which are pressed against opposite surfaces of a photographic film. From the character and size of the pattern which appears on the film after development, the polarity and the peak value of the voltage surge may be estimated, and, therefore, if the instrument is being used for current measurement, the peak value of the current can be determined with the additional knowledge of the value of the resistance shunt. These instruments have been built with several elements for recording a number of surge voltages simultaneously, and a moving film has been introduced to enable successive voltage surges to be measured as a function of time. Representative records are shown in figure 24* (Perry, Webster and Baguley, 1942). With the klydonograph it is considered that the peak value of a current surge may be determined within an accuracy of about 25 to 50 per cent.

A further simple method for the measurement of the peak current and the polarity of lightning discharges has been developed by Foust and Kuehni (1932) and Foust and Gardner (1934). Its essential feature is a "magnetic link", which consists of a small number of short steel strips of high retentivity. The link is mounted at a known distance from a conductor through which the lightning current is expected to pass. Experiments in the laboratory show that the remanent magnetism of such a link after the passage through the conductor of a current surge, which may rise to its crest value in 0·5 microsecond and fall to half-value in 5 microseconds, is the same as that caused by a direct current of the same peak value. Measurement of the remanence possessed by a link after a lightning discharge then enables a determination of the peak value of the current surge carried by the neighbouring conductor. For accurate measurement it is usual to mount two or more links at different distances from the same conductor (Foust and Henderson, 1935). Because of relative simplicity extensive use has been made of magnetic links in lightning-current measurements (Zaduk, 1935; Grünewald, 1935; Berger, 1936; Bell, 1940; Hansson and Waldorf, 1944; Gross and Lippert, 1945; Lewis and Foust, 1945).

While the magnetic link enables the current in a discharge to be determined, it gives no information concerning the variation with time of the current pulse and does not distinguish between the various strokes of a multiple discharge. In order to determine these additional facts, Wagner and McCann (1940) have developed an instrument called the fulchronograph. This consists of an aluminium wheel round the rim of which are slots containing magnetic links of sufficient length to project on both sides of the wheel. As the wheel is rotated, the links successively pass through a pair of narrow coils through which flows the current to be measured. The current at the instant during which a particular link traverses the coil can be determined by a subsequent measurement of the residual flux in the link, and, therefore, a curve relating the variation of current with time can be obtained. The time scale is governed by the speed of rotation of the wheel and, on account of the big variations in the duration of lightning discharges, two fulchronographs with different rotational speeds are usually employed at a single recording station. A considerable amount of useful data has been collected with these relatively simple instruments, which have been installed on a number of high buildings and transmission lines (Wagner, McCann and Beck, 1941).

* Facing page 331.

The cathode-ray oscillograph gives the most satisfactory record of the magnitude of the lightning-discharge current and its variation with time, but the elaborate nature of the equipment has limited its use, and it has been installed only at recording stations where direct strokes are most probable. Because of its high writing speed, the continuously evacuated type of instrument (Lee, 1928; Norinder 1928; Berger, 1928; Torok and Fielder, 1929; Ackermann, 1930; Perry and Jobling, 1936; Stekolnikov, 1937 a) has been generally used in studies of the voltage and current surges produced by lightning discharges (Fortescue, Atherton and Cox, 1929; Smeloff and Price, 1930; Sporn and Lloyd, 1930; Conwell and Fortescue, 1930; Cox and Beck, 1930; Pittman and Torok, 1931; Berger, 1934; Stekolnikov and Valeev, 1937; Perry, 1941), but it has the disadvantage that it requires constant supervision. Sealed-off cathode-ray tubes (Kuehni and Ramo, 1937) have, however, been developed with a writing speed sufficient to record surges of short duration, and these have been used in certain investigations. A limitation of the oscillograph is that, if the time resolution adopted is high enough to show details of the short-duration surges, the recording conditions are not suitable for the recording of the currents of long duration which may be present in the multiple-stroke discharge. By the use of two oscillographs this difficulty may be overcome, and for the slow-time sweep records a crater-lamp magnetic type of oscillograph is employed (McMorris, Rusher and Hagenguth, 1934).

In his observations of lightning discharges to the Empire State Building, New York, McEachron (1939) used an automatic sealed-off cathode-ray oscillograph for recording the short-time high-peak-current pulses and a crater-lamp oscillograph for the long-time low-amplitude currents. Details of these instruments are given by Hagenguth (1940). The cathode-ray oscillograph was designed to give records of surges lasting up to 100 microseconds, with the time-sweep control-circuits arranged so that subsequent surges could be recorded at intervals of 500 microseconds. A resistance shunt, consisting of linear and non-linear resistance-units connected in parallel, enabled measurements to be made of current values between 10,000 and 200,000 amp. with fair accuracy. The flourescent screen of the oscillograph was photographed on a moving film. In the crater-lamp oscillograph used by McEachron the light source was a neon-crater lamp in which the light beam becomes fully established within a time of 20 microseconds after the application of an impulse voltage to the lamp. Permanent-magnet galvanometers are used for producing the light-beam deflections. The recording film is carried on a rotating drum which enables a resolution of 440 microseconds per mm. to be obtained. The resistance shunt used with this instrument is again a combination of linear and non-linear resistance units designed so that currents from 50 amp. upwards can be measured.

Both the above instruments were installed in the dome of the Empire State Building in order that measurements might be made of the currents caused by lightning discharges when these strike the short-wave antenna which surmounts the dome. The resistance shunts were mounted at a nodal point of the antenna, so that they caused no interference with broadcasting. For large currents, flashover of the insulation of the antenna sometimes occurred, so that a part of the

lightning current flowed to the earthed structure of the building through the framework supporting the antenna. The conductors beneath the framework were equipped, therefore, with magnetic links to measure any such overflow current.

Stekolnikov and Valeev (1937) have used a cathode-ray oscillograph to record the lightning currents flowing to earth through an aerial about 1,000 m. long suspended from a balloon. The oscillograph, which was of the continuously-evacuated type with a cold cathode, possessed an arrangement whereby the cathode beam was generated only when the discharge current occurred, so that the usual beam-trapping circuits were dispensed with. Simultaneous current-measurements on the same aerial were made with magnetic links.* Klydonographs with high-speed moving films were also used, connected to antennae suspended horizontally above the ground. The resolution obtained in the latter records was sufficient to enable a distinction to be made between the patterns produced by the separate strokes of a multiple discharge.

A further instrument, which records the variations of current with time, has been developed by McCann (1942 a–44). A photograph is obtained, on a moving film, of the luminosity produced by the flow of a surge current across a short spark-gap, and it is claimed that with a calibrated instrument of this type the current magnitude can be determined to within a factor of two over the wide range of currents between 0·1 and 150,000 amp.

Indirect estimates of the characteristics of lightning discharges can be deduced approximately from the damage caused to solid materials which are struck by lightning. The tips of lightning rods may be melted; wires may be fused or become welded together; thin metal surfaces may be punctured or blistered; trees and other wooden material may be splintered; glass window-panes and porcelain insulators may be punctured or shattered; sand may be fused into fulgurites. From the examination of these effects, in comparison with the results of experiments on similar objects when subjected to large current-surges in the laboratory, various investigators have made deductions concerning the magnitude of the current and the charge conveyed to earth by lightning discharges (McEachron and Thomason, 1935; Bellaschi, 1935–41; Cozzens, 1939; McEachron and Hagenguth, 1942). Laboratory studies by Bellaschi (1941) and McEachron and Hagenguth (1942) show that the amount of metal fused is a function primarily of the total charge conveyed, and estimates of the latter can be made, therefore, by examination of metal objects which have been struck by lightning. The explosive effects caused by lightning result from high-peak currents of short duration; the burning effects are produced by currents of relatively low values, of the order of a hundred ampères, lasting for times about 0·1 sec. or more. The large magnetic forces accompanying high-peak currents may be sufficient to crush thin-walled conductors, as shown in figure 25†, where an unused tubular lightning-rod is compared with a similar rod which has been subjected to a lightning discharge (Humphreys, 1940).

* A brief description of measurements with magnetic links of the current pulses produced in the cables of barrage balloons struck by lightning has been given recently by Davis (1945)

† Facing page 331.

(ii) *The magnitude of lightning currents*

Because of the complex nature of the lightning discharge and the limitations of the various measuring devices used to measure the stroke currents, it is certain that most of the results are to some extent incomplete and inaccurate. Sufficient data have been collected, however, to show that there is a wide range of currents involved, and the main problem is to obtain a correct statistical analysis of the frequency of distribution of the current magnitudes and wave-shapes.

The frequency of occurrence of lightning discharges as a function of their peak current is shown in figure 26. A peak current of 160,000 amp. has been recorded by a fulchronograph (Wagner, McCann and Beck, 1941), and a still larger current of 220,000 amp. has been deduced from records given by magnetic links installed on transmission-line towers. The chances of such large currents

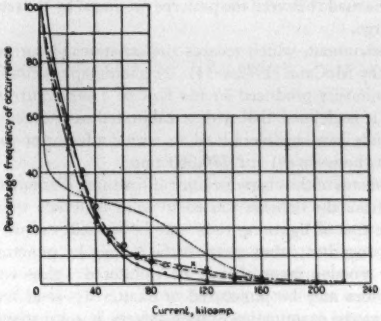


Figure 26. Maximum currents observed in lightning discharges (Bruce and Golde, 1942).
(Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.)

● · - · - · - · ● Grunewald.
⊙ - - - - - ⊙ Lewis and Foust.
× — · — · — × Stekolnikov and Lamdon.
. . . . . McEachron.

The points represent the proportion of current values exceeding that shown on the abscissa-axis.

are remote, and a peak value of about 20,000 amp. is the average observed. These peak currents are of short duration, as shown in the oscillograms given in figure 27. Some investigators (Stekolnikov and Valeev, 1938; Schonland, Malan and Collens, 1935) have found that the first stroke of the discharge is usually strongest and that subsequent strokes show a general tendency to become progressively weaker, though individual strokes in a series may contradict the general trend. On the other hand McEachron (1939) has recorded multiple-stroke discharges to the Empire State Building in which the largest peak-current value occurs after a number of preceding weaker strokes, and a similar result is shown in certain fulchronograph records (Wagner, McCann and Beck, 1941).

The question of the polarity of the discharge current has long been a subject of controversy and has been linked, naturally, with the problem of the distribution of charges in the thundercloud. The measurement of the discharge current by direct and indirect methods has to some extent clarified the issue, and it is now known that currents conveying negative charge to ground predominate. According to Bruce and Golde (1942) a distinction is observed between discharges in the tropics and those in temperate zones. For the former it appears that the number

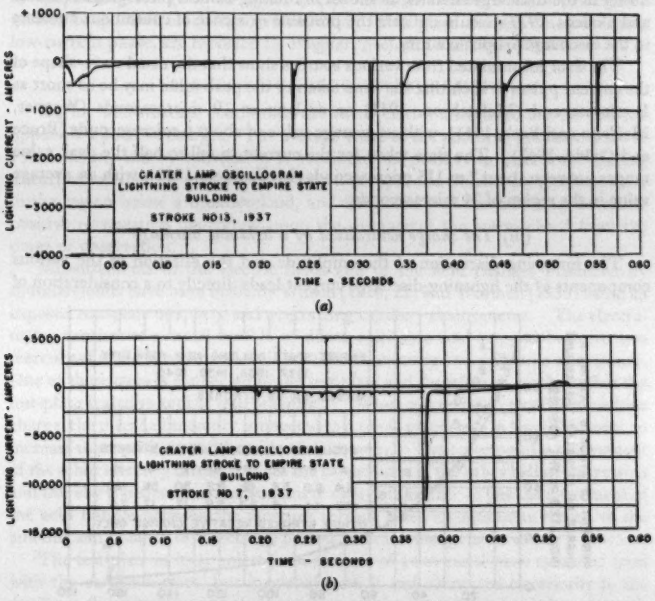


Figure 27. Oscillograms, of the variations of current with time in two separate lightning discharges, as recorded with a crater-lamp oscillograph (McEachron, 1939).

(*Reproduced, by kind permission, from the 'Journal of the Franklin Institute'.*)

of negative discharges exceeds the number of positive discharges in the ratio of 17 to 1, whereas in temperate zones the ratio is only 3 to 1. Out of 49 lightning discharges to the Empire State Building, McEachron (1941) observed positive currents in only eight cases. The positive current-pulse is invariably preceded or followed by negative current-pulses during the same discharge.

In addition to the short-time current-pulses, low-amplitude currents of long duration are usually observed in the lightning discharge. These currents range in value from tens of ampères to a few thousand ampères, and the term "continuing

stroke" (McEachron, 1939) has been used to denote their presence. In some cases the continuing stroke persists for the duration of the discharge, so that current flow is maintained at a comparatively low level, with superimposed high current-pulses. In other cases the current-flow ceases and is followed by isolated current-pulses of higher magnitude. These observations were made for discharges to a high structure, but Wagner, McCann and Beck (1941) have shown that similar results may be obtained for discharges to transmission lines, and the persistence of luminosity in the discharge channel, as shown in rotating-camera photographs (Malan and Collens, 1937), would indicate the probable existence of continuous currents in the discharge to open country.

The data accumulated from various sources show that the usual wave-shape of the current pulses is such that the time taken to the peak value may be as short as 1 microsecond (McEachron, 1941) or as long as 19 microseconds (Wagner, McCann and Beck, 1941), with an average value of about 6 microseconds (Bruce and Golde, 1942). The time taken for the current to fall to half the peak value ranges between about 7 to 115 microseconds (McEachron, 1941), with an average value in the region of 24 microseconds.

(iii) *The charge neutralized by a lightning discharge*

The foregoing discussion of the amplitude and the duration of the various components of the lightning-discharge current leads directly to a consideration of

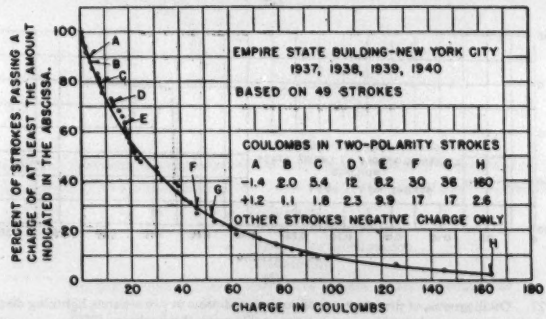


Figure 28. Total charge neutralized by lightning discharges (Wagner and McCann, 1941). (*Reproduced, by kind permission, from 'Electrical Engineering'.*)

the amount of charge neutralized by each of these components, and thence directly to the total charge neutralized in a multiple-stroke discharge. The short-time pulses, of large peak current, generally carry only a small proportion of the total charge conveyed to earth. In measurements by McEachron (1941) the charge for each current-pulse varies from less than 0·1 coulomb up to 5 coulombs. Only 50 per cent of the individual strokes in a discharge are associated with a charge of 0·13 coulomb or more, and only 5 per cent with more than 1·5 coulombs. On the

other hand the analysis given in figure 28 of the total charge involved shows that at least 50 per cent of the discharges convey a charge of 25 coulombs or more. In several of the discharges a reversal of current flow took place, and an analysis of the amounts of charge of both polarities which are conveyed is included in figure 28, where it will be observed that the negative charge predominates. The maximum charge observed in McEachron's observations was 164 coulombs, of which by far the greater proportion was conveyed to earth by a continuing low current, as shown in the oscillogram given in figure 27 (*b*). Similar observations, in which the greater proportion of charge neutralization takes place during the continuing low-current phase, are reported by Wagner, McCann and Beck (1941).

§5. ELECTRIC FIELD MEASUREMENTS

During fine-weather conditions, the earth carries a negative charge and the upper air a positive charge. The average value of the potential gradient is about 100 volts/m., and by convention is considered as positive in direction. During disturbed weather the gradient is subjected to rapid and violent changes, especially in the region below a thundercloud, and the polarity of the field may be either positive or negative, depending upon the distance of the storm cloud from the point of observation.

Measurements of the electric fields and of the field changes produced by thunderclouds have been made by Wilson (1916, 21) and Wormell (1939) using an exposed test-plate or sphere and a recording capillary electrometer. The electrometer consists of a small bubble of dilute sulphuric acid trapped between two mercury threads in a narrow capillary tube connecting two end-cups of mercury. One of these cups is connected to the test plate and the other to earth, so that the test-plate is always kept at zero potential. When one end-cup receives a positive charge, the thread of mercury connected to it tends to increase in length in order to increase the capacitance of that side of the system. The corresponding movement of the other mercury thread reduces the capacitance of the other half of the system and thereby transfers an equal amount of charge to earth. Thus the movement of the acid bubble between the mercury threads gives a continuous record of the amount and polarity of the charge passed from the test-plate to earth.

The test-plate in its simplest form consists of a flat metal plate mounted level with the earth's surface, but insulated from it and connected electrically to the capillary electrometer. At the beginning of the recording an earthed cover is placed over the test-plate and the electrometer bubble assumes the zero position. When the earthed cover is removed, a charge due to the earth's field is induced on the test-plate and the electrometer bubble moves rapidly (within 0·05 sec.) to the equilibrium position for which this charge is just compensated. Thereafter the bubble movements follow the variations of charge on the test-plate, due either to a change in the electric field, to conduction current, to point discharges or to transfer of charge by falling rain. The earthed cover is swung over the test-plate at frequent intervals in order to determine the zero point from which the field and field changes can be measured. The movement of the electrometer bubble bears a linear relationship to the charge passed through the instrument, and hence the

induced charge Q on the test-plate can be determined at any instant. If A is the area of the test-plate, then F, the gradient, is given by

$$F = -\frac{4\pi Q}{A}. \qquad \ldots\ldots(2)$$

In addition, the charge passed in a given time, due either to conduction current or to the falling of charged rain, can be determined from the records of the electrometer movement.

In order to obtain a continuous record of the field conditions and changes, the acid bubble is illuminated from below and a highly magnified image of the bubble-edge is projected (by means of a microscope) through a fine slit on to a photographic plate which is moved horizontally at uniform speed in a direction at right angles to the slit.

Figure 29 (Wormell, 1939) shows a typical record produced during a storm with the sudden changes of field caused by distant lightning-flashes. The vertical black lines on the record are due to the illumination of the electrometer bubble being interrupted at the instant that thunder is heard. The interval between the sudden field-change due to the flash and the mark made when thunder is heard enables an estimate to be made of the distance of the flash from the point of observation. At the point marked X near the right-hand side of the record, the test-plate was covered by the earthed plate to determine the zero reference line. The record shows six positive field changes of the order of 1,000 volts/m., followed in each case by a comparatively slow recovery of the field to its former value. This recovery process gives an exponential type of curve on the record, and as Wilson (1921) has pointed out, this probably means that when the flash occurs the separated charges in the thundercloud are destroyed, but thereafter the charges begin to be re-established at a rate proportional to the initial rate of recovery of the field. As the separated charges increase, the field due to them tends to diminish the rate of increase in charge until finally a state of equilibrium is reached. In certain cases, however, the initial rate of recovery of the field is very small but increases later.

In considering the analysis of the field records and the changes of field observed when lightning flashes occurred, Wilson (1921) determined theoretically the effect of lowering a charge Q from a height H_2 to a height H_1. At ground level the sudden change of field, associated with the lowering of a positive charge in this way, would be positive if the point of observation were less than a certain critical distance from the charge centre and negative for greater distances. At the critical or reversal distance, zero field-change would be expected, though in practice, owing to the fact that the thundercloud charges may occupy a large area, the reversal distance should be considered rather as a reversal zone where small field-changes of either sign may be expected. Hence for a bi-polar cloud with the upper portion charged positively and the lower portion charged negatively, when the point of observation is near to the cloud the field will be negative just before a flash and will show a positive field-change when the cloud discharges within itself. Beyond the critical distance, however, the field will be initially positive, but will show a negative field-change on the occasion of a lightning discharge. Similar measurements (Schonland and Craib, 1927; Schonland, 1928 a) in South Africa showed that beyond 15 km.

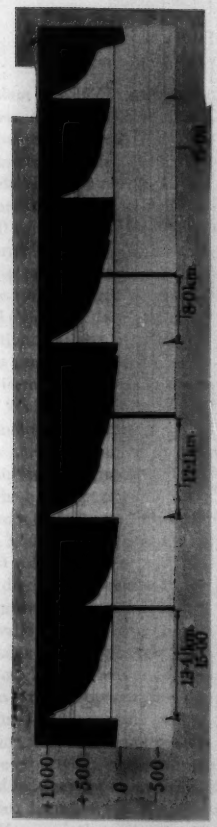


Figure 29. Representative record of field changes detected with a capillary electrometer (Wormell, 1939).

(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)

To face page 344

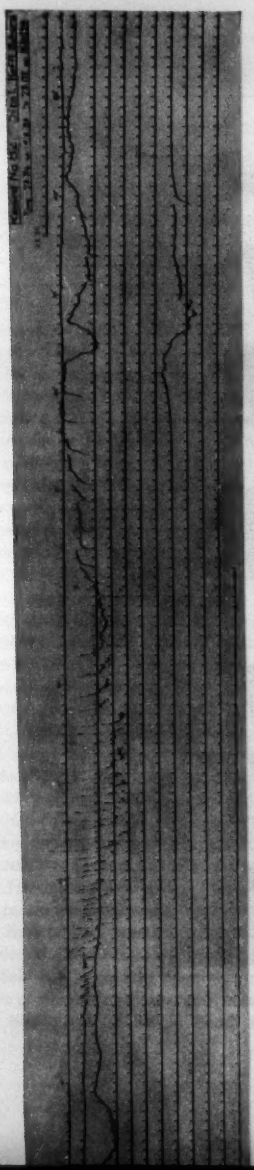


Figure 30. Typical record of variations in point-discharge current (Perry, Webster and Baguley, 1942).
(*Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.*)

negative field-changes recorded were 2,375 in number, compared with only 250 positive field-changes; for distances less than 7 km. positive field-changes were 188 against 9 negative field-changes. Wilson (1921) showed that when L, the horizontal distance between the point of observation and the storm-cloud, is large, then F, the vertical electric force on a point on the ground due to the lowering of a charge Q from a height H_2 to a height H_1, is approximately given by

$$F = 2Q(H_2 - H_1)/L^3,$$

so that

$$FL^3 = 2Q(H_2 - H_1). \qquad \ldots\ldots(3)$$

Therefore, for large values of L, FL^3 becomes equal to the electric moment of the discharge. From these and other considerations Wilson deduced that the average value of the charge in a cloud, destroyed by the lightning flash, is of the order of 20 coulombs, but later workers (Bruce and Golde, 1942) considered that an average value of 30 coulombs gives a better interpretation of Wilson's results.

The value of the pre-discharge potential-gradient below a storm is found to range from 0 to 10,000 volts/m. This latter value, though not the maximum, is rarely exceeded (Wormell, 1939). The values of the sudden field-changes caused by lightning discharges decrease with increasing distance of the point of measurement from the storm centre. For distances of within 2 km. individual field-changes observed (Wormell, 1939) have values varying from 1,000 to 61,000 volts/m. The resolution of these measurements is insufficient to distinguish between field-changes caused by individual strokes of a discharge. Some of the more complicated forms of field-change are referred to in the discussion which follows on the work of Whipple and Scrase (1936).

The electric field and its variations during disturbed weather conditions can be studied, indirectly, by the behaviour of the discharge current from an elevated point on the earth's surface. Schonland (1928 b) found in South Africa that a small tree 4 m. high which was cut off at its base and mounted on insulators gave a current of 0·07 microamp. (as measured by a galvanometer connected to earth) in a field of gradient $-3{,}500$ volts/m. and a current of 4·0 microamp. when the gradient was $-16{,}000$ volts/m. Wormell (1927, 29) at Cambridge studied the quantity of electricity discharged, positive and negative, from an elevated point by means of a water micro-voltmeter connected between the point and earth. In three years of observations the average quantity of excess positive electricity discharged from the point to the air per year was approximately 0·13 coulomb. In analysing the conditions of 147 periods of disturbed weather, accompanied by rain, 103 periods showed an excess of upward-flowing current (i.e. positive discharge—or outflow), 34 an excess of downward current (negative discharge—or inflow), whilst in the remaining 10 periods the net discharge was zero. These values agreed with those to be expected in view of the predominance of negative potential gradients in disturbed weather conditions. A more detailed study of the point-discharge current during disturbed weather conditions was made at Kew (Whipple and Scrase, 1936). Here a platinum point raised to a height of 8·4 m. above the ground was connected through a critically-damped galvanometer to earth. The galvanometer deflections were recorded photographically on sensitive paper carried

on a drum which revolved once every two hours. The drum was fitted with an axial traversing mechanism to avoid overlapping of the two-hourly records, and continuous recording throughout 24 hours was possible before the drum had to be reloaded. During the same period, continuous records were made of the potential gradient by means of a polonium collector connected to an insensitive oil-damped Dolezalek electrometer. By comparing the two sets of records it was possible to deduce a relationship between the point-discharge current and the potential gradient such that

$$i = a(F^2 - M^2), \qquad \ldots\ldots(4)$$

where i = point-discharge current in microampères,

a = constant ($a = 0{\cdot}0008$ for positive gradients and $0{\cdot}0010$ for negative gradients);

M = estimated mimimum field strength at which discharge takes place ($M = 7{\cdot}8$ volts/cm. for positive gradients and $8{\cdot}6$ volts/cm. for negative gradients);

F = gradient giving a discharge current i.

The highest current recorded during the period of the observations (August 1932 to December 1934), excluding the large surge-currents due to lightning, was -16 microampères, which corresponded to a gradient of about $-12{,}600$ volts/m.

The type of record given by the point-discharge recorder is illustrated by figure 30, which shows the variations in point-discharge current during a storm in Nigeria (Perry, Webster and Baguley, 1942), where a similar investigation to that at Kew was made in 1939. The point-discharge current was negligibly small until 01.20 h., when it was just perceptible on the record. It became a large outflow current (from point to atmosphere corresponding to a negative gradient) soon after 02.20 h., when the storm cloud was nearly overhead. Discontinuities showing high-peak inflow currents occurred roughly at minute intervals, corresponding to lightning flashes from the storm cloud, but at 03.00 h. the steady current became an inflow current (positive gradient) with high-peak outflow-currents superimposed at less frequent intervals. After further changes in polarity and amplitude the discharge current ceased (at about 04.30 h.).

The measurements of the point-discharge currents at Kew, Cambridge and Nigeria, discussed above, all show very similar characteristics, though there were small differences, due to local conditions of exposure and to frequency of storms. At Cambridge the average yearly excess of outflow charge over inflow charge (for two years of observation) was 130 millicoulombs, with a ratio of outflow of 2·0 to 1. At Kew the corresponding figures for a similar period were 60 millicoulombs and ratio of 1·7/1. In Nigeria, for about three-quarters of one lightning season, the figures were 361 millicoulombs and a ratio of 2·68/1. At Kew the diurnal variations in discharge were analysed, and for the whole year showed a maximum discharge in the afternoon, with little activity between 20 h. in the evening and 10 h. in the morning. The discharge was considerably greater in the summer and equinox than in the winter. A comparison of the diurnal variation (for two years' observations) of the net point-discharge with the diurnal variations of thunderstorm activity showed a striking resemblance between the two curves, and this was even

more marked by the Nigerian results. Figure 31 shows a composite picture of the Kew and Nigerian observations linked with thunderstorm activity over the whole globe. The reason for this is that, for thunderstorms over the land, the activity is most marked in the later afternoon, and most of the large land masses (Central Africa and South America), where storms are particularly active, are close to the longitude of Greenwich (and Nigeria). Hence the variations in point discharge are shown to be linked with thunderstorm activity, and Wilson's (1921) contention that the point discharge is an important factor in the exchange of electricity between the earth and the atmosphere is well supported by all these observations. A comparison between the inflow of positive electricity to the earth due to the fine-weather current and precipitation of rain, and the outflow due to lightning and point discharge, shows that the latter two phenomena are probably

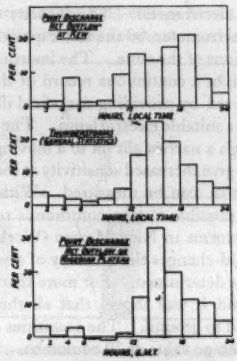


Figure 31. Comparison of diurnal variations of net point-discharge current, for Kew and Nigeria, with thunderstorm activity over the globe.

(*Reproduced, by kind permission, from the 'Journal of the Institution of Electrical Engineers'.*)

more than sufficient to maintain the earth at a negative potential with respect to the upper atmosphere (Wormell, 1930; Whipple and Scrase, 1936).

The analysis of the point-discharge records at Kew (and in Nigeria) show that the field-changes causing them are of a varied character. When the preliminary field was positive, three types of field-change were identified, namely: pP—positive preliminary field followed by a positive field-change; pND—positive preliminary field followed by a negative field-change, which decreases the positive field; and *pNR*—positive preliminary field followed by a negative field-change sufficiently large to cause a reversal in field polarity. With preliminary fields of negative polarity, the changes were again of three kinds; nN, nPD, and PR, which are the complementary changes to those defined above. Two other types of field-change were classified for those cases where the preliminary field was small and indeterminate, namely, that where the field-change was positive (SP) and the

other where the field-change was negative (SN). All these changes are similar to those recorded by Wilson (1921) and Wormell (1939). Whipple and Scrase (1936) show that the presence of a space-charge blanket between the ground and the thundercloud would have an important effect in the vertical distribution of field below the cloud and in the field-changes occurring immediately after a lightning discharge. The presence of such a space charge has, however, yet to be confirmed.

In North America the study of the electric fields at the surface of the earth during thunderstorm activity have been studied by Workman and Holzer (1939 a) using a group of generating voltmeters. In this type of instrument, a metal vane, shaped like a double sector, is rotated under an earthed shield containing two apertures. When the rotating vane is below the openings it is earthed, but acquires an induced change of magnitude proportional to the existing field. As the vane moves to the shielded position it is connected by means of a commutator to the needle of a quartz electrometer. The capacity of the metal vane is much larger than that of the electrometer, so the latter quickly attains the full potential within a very few revolutions of the vane. The instrument has a resolving power of less than 1 sec. and gives a continuous record of the electrical field whilst in operation. This is achieved by providing dark-field illumination for the electrometer needle by means of a suitable electric lamp. The movement of the needle is projected by a lens through a narrow slit on to a moving film. The electrometer plates are so shaped as to give increased sensitivity at the lower field-strengths, but fields up to 100,000 volts/m. can be measured. With five such instruments at different locations it was possible to take continuous records of the electric-field intensity of a number of storms in New Mexico (Workman and Holzer, 1939 b).

From the recorded field-changes the quantity of charge delivered to the earth during thunderstorms was determined. For more than 200 discharges occurring between cloud and ground it was shown that all these discharges transferred negative charge from cloud to ground. The quantities involved ranged from 5 to 200 coulombs, with an average value of 30 coulombs.

The investigations described above have been concerned with the electrostatic field and its changes in the region of thunderclouds. The change of field occasioned by an electrical discharge is, however, the sum of three components, as given in the following equation:

$$E = \frac{M}{r^3} + \frac{1}{cr^2}\cdot\frac{dM}{dt} + \frac{1}{c^2 r}\cdot\frac{d^2M}{dt^2}, \qquad \ldots\ldots(5)$$

where E = instantaneous value of the field, in e. s. u.,

r = the distance of the observation point from the discharge;

M = electric moment of the composite dipole formed by the electric charges and their images in the ground, in e. s. u.;

c = velocity of light.

The first term on the right-hand side of the equation represents the electrostatic field, the second term the magnetic field, and the third term the radiation field. The first term is of most importance in the neighbourhood of the discharge, and the third term predominates at considerable distances away from it.

The variations of the electrostatic field-changes close to a discharge have been recorded, using an oscillograph, by Appleton and Chapman (1937) and Schonland Hodges and Collens (1938). In the latter measurements the output from an aerial is transferred to a cathode-ray oscillograph, whose screen is photographed with a film which is carried on a rotating drum provided with an axial traversing motion (Lutkin, 1937). The apparatus is essentially similar to that used in the observation of atmospherics by Watson Watt, Herd and Lutkin (1937). A typical form of oscillogram is shown diagrammatically in figure 32, where it is related to the photographed image of the leader- and main-stroke sequence (Schonland, Hodges and Collens, 1938). Three main changes in the field occur, as was shown first by Appleton and Chapman (1937), who identified the "*a*" portion with the leader-stroke, the "*b*" portion with the return stroke and the "*c*" portion with the continuation of the discharge after the return stroke has reached the cloud. Oscillations on the "*a*" portion of the field-change have a period corresponding to that observed between the individual steps of the stepped leader-stroke.

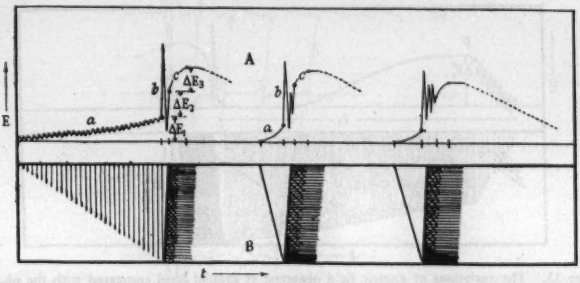


Figure 32. The variations of electric field observed at ground level compared with the photographed development of the associated lightning discharge, as observed in about 65 per cent of discharges to ground in South Africa (Schonland, Hodges and Collens, 1938).

(*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

The electric-field variations illustrated in figure 32 are observed in about two-thirds of the discharges recorded by Schonland, Hodges and Collens (1938). In most of the other discharges the field varies in the manner shown in figure 33, where again it has been possible to correlate the electrical characteristics with the photographed development of the discharge.

The total magnetic field-strength at a distance r from the electrical discharge is given by

$$H = \frac{1}{cr^2}\frac{dM}{dt} + \frac{1}{c^2 r}\frac{d^2M}{dt^2}, \qquad \ldots\ldots(6)$$

where H is measured in electromagnetic units and all the other terms have the same significance as in equation (5). Norinder (1935, 37) has recorded by means of an oscillograph the magnetic field-changes produced by lightning discharges, and from his data has estimated the magnitude of the currents in the discharge. As shown by Bruce and Golde (1942), certain of Norinder's interpretations are

misleading and cannot readily be reconciled with the findings of other workers. The values of the amplitude of the magnetic field-changes for lightning discharges a few kilometres away range from 0 to 5×10^{-3} e.m. units (Norinder, 1937).

Field-changes of the type given by the third term of equation (5) are observed as "atmospherics" in radio receivers and other communication circuits, and it is now established that quite a high proportion of atmospherics, if not all, originate in lightning discharges occurring at some part of the earth (Schonland and Hodges, 1936). By means of the direction-finding cathode-ray oscillograph equipment developed by Watson Watt, Herd and Bainbridge-Bell (1933) it is possible, by means of two frame aerials (one on the N–S axis and the other on the E–W axis) coupled through amplifiers to the X and Y plates respectively of a cathode-ray oscillograph, to determine both the magnitude and the direction of origin of any disturbance of sufficient magnitude to be recorded on the instrument. In South Africa such an investigation (Schonland and Hodges, 1936) was combined with meteorological observations on the land mass and to a less extent on the oceans to

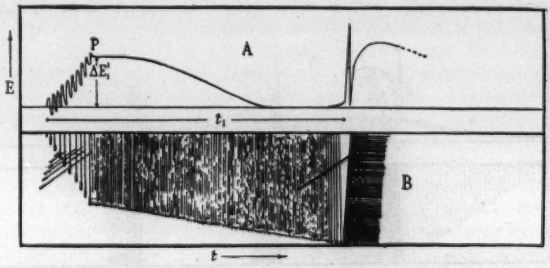


Figure 33. The variations of electric field observed at ground level compared with the photographed development of the associated lightning discharge, as observed in about 35 per cent of discharges to ground in South Africa (Schonland, Hodges and Collens, 1938).
(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)

east and west. This investigation showed that the relation between recorded atmospherics and known storms was sufficiently close to suggest that, with more complete meteorological data, a complete link between the two phenomena might have been demonstrated. The use of a single direction-finding equipment can give an approximate distance for the storm by observing the mean amplitude of the atmospherics radiated by it. A more accurate determination of storm location can be made by triangulation from two or more equipments recording the same atmospheric disturbances. By this means the distance to the source of atmospherics can be located to an accuracy of the order of 10 per cent (Lutkin, 1938, 41).

A large amount of attention has been paid to the recording of the wave-shapes of atmospherics and to correlating their forms with the characteristics of the lightning flash. At long distances (up to 500 km.) the wave-shape of atmospherics recorded by Schonland, Elder, Hodges, Phillips and van Wyk (1940) shows a simple pulse followed by similar pulses, usually of smaller amplitude, and occurring at increasing intervals of time. According to these observers the first pulse is the

ground pulse G, having characteristics which are predictable from the known behaviour of the lightning discharge, followed by a series of sky pulses, S_1 to S_n, produced by successive reflections between the ionosphere and earth. Figure 34 shows typical atmospheric wave-forms recorded at various distances from the source of origin, in which the various pulses are separated by quiet intervals.

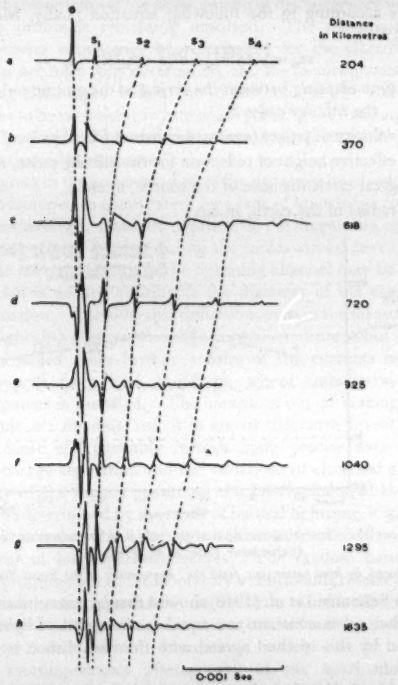


Figure 34. Typical atmospheric wave forms observed at different distances from a lightning discharge (Schonland, Elder *et al.*, 1940).

(*Reproduced, by kind permission, from the ' Proceedings of the Royal Society '.*)

Other more complicated forms are observed in which the length of the pulses is comparable with the time between successive pulses in the early part of the series, so that successive pulses tend to merge into one another without a clear separation.

Wave-forms were recorded simultaneously, at night, in Johannesburg and Durban (*g* and *h* of figure 34 originate from the same lightning flash, located at 1,295 km. from Johannesburg and 1,635 km. from Durban). In addition,

direction-finding equipments at these two stations were operated simultaneously, prior to recording wave-shapes, in order to locate the sources of all atmospheric disturbances occurring at the time of the observations. The direction-finding equipment was therefore used as a check to determine the distance of origin of disturbances, whilst an analysis of the wave-forms of these disturbances was made to give an independent measurement of this distance and also of the height of the ionosphere according to the following equation (Laby, McNeill, Nicholls and Nickson, 1940):

$$ct_n = \sqrt{4n^2h^2{}_n + d^2(1 + h_n/R)} - d, \qquad \ldots\ldots(7)$$

where t_n = time elapsing between the arrival of the ground pulse G and that of the nth sky pulse S_n;

c = velocity of pulses (assumed constant for all pulses), in cm. per sec.;

h_n = effective height of reflection for the nth sky pulse, in cm.;

d = great circle distance of the source, in cm.;

R = radius of the earth, in cm.

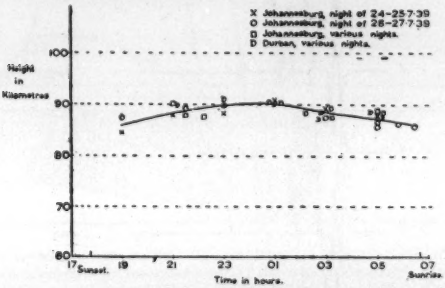


Figure 35. Nocturnal variation of the height of the reflecting layer (Schonland, Elder *et al.*, 1940).

(Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.)

The analysis by Schonland *et al.* (1940) showed that h_n was substantially the same for all pulses, that c was constant and equal to the speed of light, and that the value of d found by this method agreed with that established by the direction-finding equipment.

The value of h_n at night was found to be of the order of 90 km., with slightly smaller values near sunset and sunrise, as shown by figure 35. Some measurements of the height of the ionosphere during the day gave a value of 60 km. (Schonland, Elder, van Wyk and Cruickshank, 1939; Schonland, Elder, Hodges, Phillips and van Wyk, 1940).

Although the analysis of atmospheric wave-forms by night seems to be satisfactorily explained by the above analysis, other explanations have been offered for the wave-form of atmospherics by day. Lutkin (1939) suggests that the daylight wave-form is due to oscillations and multiple discharges in the lightning channel. Laby, McNeill, Nicholls and Nickson (1940), although originally suggesting the

reflection phenomenon, now consider that the first few half-periods of the daylight atmospheric wave-forms arise directly from the characteristics of the lightning discharge itself.

§6. FUTURE WORK

Although there is a wealth of information concerning the lightning discharge, and the general characteristics of the development of discharges are well established, there are many problems remaining unsolved. The relative importance of the various processes which may be responsible for the electrification of the thundercloud has not been fully determined, and the factors governing the initiation of lightning discharges within the cloud still require examination. As yet, there appears to be no satisfactory theory explaining the discharge mechanism, and proof is still required of some of the various hypotheses which have been put forward, such as the " pilot streamer ".

Further progress in the understanding of the processes involved in the lightning discharge is handicapped to some extent by a lack of knowledge concerning some of the parameters involved. Data are required for the magnitude, and the variation with time, of the currents flowing during the leader-stroke development. Also, in order that the current density in the lightning channel may be assessed, additional information is needed relating to the diameter of the channel at various stages in its formation. Considerable difficulties exist in the measurement of these quantities for lightning discharges under normal conditions, but useful information may be obtained from further studies of the currents in leader-strokes developing upwards from high buildings, where cathode-ray oscillographic recording-equipment is installed. The complications of making measurements in thunderclouds are obvious, but it is hoped that such investigations will be continued, as these will probably furnish more precise data concerning the mechanism of charge separation and the variations of electrical gradient.

While many of the factors governing the development of the lightning discharge have to be determined by the study of natural lightning, it is clear that much may be learnt concerning the lightning mechanism from similar investigations of the development of long spark-discharges. The random nature of lightning impedes rapid progress, but in the laboratory a continuous research programme on the spark mechanism can be pursued. In the latter case the different variables can be controlled more readily and measurements can be made by means of an oscillograph of voltage and current fluctuations which can be compared with synchronized rotating-camera photographs of the spark development. At present, although the spark mechanism in nearly uniform fields has been explained theoretically, there is no satisfactory explanation of the spark mechanism in long non-uniform gaps, such as exist between a point and a plane. Once a theoretical explanation can be given, a considerable step forward will have been made in the fuller interpretation of the lightning mechanism. For this reason the study of the spark in the laboratory is a desirable adjunct to the study of natural lightning.

Besides the investigation of the physical mechanism of lightning, there is a need for further statistical information concerning such quantities as the maximum current in a discharge, the rate of rise of current, the number of multiple strokes, and the total charge conveyed to ground. Much of this type of data has already

been obtained, but it is as yet insufficient to provide the satisfactory statistical analysis required by the designer of apparatus for the protection of objects against the damaging effects of lightning.

The encouraging feature of the various research investigations on the lightning discharge is that although many of them are concerned with only one or two aspects of the phenomenon, yet the main picture of the process is now gradually emerging, and the various isolated investigations can be fitted fairly satisfactorily into the general framework. Even so, progress towards an elimination of some of the present inconsistencies and gaps in the information would be more rapid if all the aspects of the lightning discharge could be studied simultaneously. This would necessitate synchronized measurements with many recording devices, designed to cover the wide range of the variables encountered in lightning discharges, for which the currents vary from several ampères to a few hundred thousand ampères, the associated field-changes vary from microvolts per cm. to several thousand volts per cm., and the time-scale requires measurements of events taking place in times varying from microseconds to seconds. Clearly such a project would involve a large expenditure, and would require a research centre situated in a region where there is a reasonable frequency of storm occurrence.

§ 7. ACKNOWLEDGMENTS

The authors express their appreciation to Sir Arthur P. M. Fleming, C.B.E., D.Eng., Director of Research and Education, and to Mr. B. G. Churcher, M.Sc., M.I.E.E., Manager of the Research Department, Metropolitan-Vickers Electrical Company Limited, for permission to publish this review.

GENERAL REFERENCES

FLEMING, J. A. (Editor), 1939. *Terrestrial Magnetism and Electricity* (New York: McGraw-Hill).
HUMPHREYS, W. J., 1940. *Physics of the Air* (New York: McGraw-Hill).
SCHONLAND, B. F. J., 1932. *Atmospheric Electricity* (London: Methuen).
STEKOLNIKOV, I. S., 1940. *Lightning* (Moscow: U.S.S.R. Academy of Publications).

REFERENCES

ACKERMANN, O., 1930. *Trans. Amer. Inst. Elect. Engrs.*, **49**, 467.
AKOPIAN, A., 1937. *Conf. int. gr. Rés. Élect.*, Paper No. 328.
ALLBRIGHT, J. C., 1937. *J. Appl. Phys.*, **8**, 313.
ALLIBONE, T. E. and MEEK, J. M., 1938 a. *Proc. Roy. Soc.*, A, **166**, 97.
ALLIBONE, T. E. and MEEK, J. M., 1938 b. *Proc. Roy. Soc.*, A, **169**, 246.
ALLIBONE, T. E. and SCHONLAND, B. F. J., 1934. *Nature, Lond.*, **134**, 736.
APPLETON, E. V. and CHAPMAN, F. W., 1937. *Proc. Roy. Soc.*, A, **158**, 1.
BEADLE, D. G., 1936. *Nature, Lond.*, **137**, 112.
BELIAKOV, A. P. and HANOV, B., 1936. *Elektrichestvo*, 57 (section 22), 20.
BELL, E., 1940. *Trans. Amer. Inst. Elect. Engrs.*, **59**, 822.
BELLASCHI, P. L., 1935. *Trans. Amer. Inst. Elect. Engrs.*, **54**, 837.
BELLASCHI, P. L., 1937. *Trans. Amer. Inst. Elect. Engrs.*, **56**, 1253.
BELLASCHI, P. L., 1941. *Trans. Amer. Inst. Elect. Engrs.*, **60**, 1248.
BERGER, K., 1928. *Bull. Ass. Suisse des Électr.*, **19**, 668.
BERGER, K., 1934. *Bull. Ass. Suisse des Électr.*, **25**, 213.
BERGER, K., 1936. *Bull. Ass. Suisse des Électr.*, **27**, 145.
BOYS, C. V., 1926. *Nature, Lond.*, **118**, 749.

BOYS, C. V., 1929. *Nature, Lond.*, **124**, 54.
BRAND, W., 1923. *Der Kugelblitz* (Hamburg: H. Grand).
BRUCE, C. E. R., 1941. *Nature, Lond.*, **147**, 805.
BRUCE, C. E. R., 1944. *Proc. Roy. Soc.*, A, **183**, 228.
BRUCE, C. E. R. and GOLDE, R. H. 1942. *J. Instn. Elect. Engrs.*, **88** (Part II), 487.
CANDLER, J. H., 1940. *J. Instn. Elect. Engrs.*, **87**, 597.
CONWELL, R. N. and FORTESCUE, C. L., 1930. *Trans. Amer. Inst. Elect. Engrs.*, **49**, 872.
COX, J. H. and BECK, E., 1930. *Trans. Amer. Inst. Elect. Engrs.*, **49**, 857.
COX, J. H. and LEGG, J. W., 1925. *Trans. Amer. Inst. Elect. Engrs.*, **44**, 857.
COX, J. H., MCAULEY, P. H. and HUGGINS, L. G., 1927. *Trans. Amer. Inst. Elect. Engrs.*, **46**, 315.
COZZENS, B., 1939. *Trans. Amer. Inst. Elect. Engrs.*, **58**, 140.
CRAGGS, J. D. and MEEK, J. M., 1946. *Proc. Roy. Soc.*, A (in press).
DAVIS, R. 1943. *The Engineer*, **180**, 435.
EVANS, E. A., 1943. *Trans. Amer. Inst. Elect. Engrs.*, **62**, 427.
FLOWERS, J. W., 1943. *Phys. Rev.*, **64**, 225.
FLOWERS, J. W., 1944. *Gen. Elect. Rev.*, **47**, 9.
FORTESCUE, C. L., ATHERTON, A. L. and COX, J. H., 1929. *Trans. Amer. Inst. Elect. Engrs.*, **48**, 449.
FOUST, C. M. and GARDNER, G. F., 1934. *Gen. Elect. Rev.*, **37**, 324.
FOUST, C. M. and HENDERSON, J. T., 1935. *Trans. Amer. Inst. Elect. Engrs.*, **54**, 373.
FOUST, C. M. and KUEHNI, H. P., 1932. *Gen. Elect. Rev.*, **35**, 644.
FRENKEL, J., 1940. *J. Exper. and Theor. Physics*, **10**, 1424.
FRITSCH, V., 1941. *Naturwissenschaften*, **29**, 377.
GOLDE, R. H., 1941. *J. Instn. Elect. Engrs.*, **88** (Part II), 67.
GOLDE, R. H., 1945. *Brit. Elect. Allied Ind. Res. Ass. Report*, No. S/T. 49, 10.
GOODLET, B. L., 1937. *J. Instn. Elect. Engrs.*, **81**, 1.
GOTT, J. P., 1935. *Proc. Roy. Soc.*, A, **151**, 665.
GROSS, I. W. and LIPPERT, G. D., 1945. *Trans. Amer. Inst. Elect. Engrs.*, **64**, 76.
GRÜNEWALD, H., 1935. *Conf. int. gr. Rés. Élect.*, Paper No. 326.
HAGENGUTH, J. H., 1940. *Gen. Elect. Rev.*, **43**, 195 and 248.
HALLIDAY, E. C., 1932. *Proc. Roy. Soc.*, A, **138**, 205.
HANSSON, E. and WALDORF, S. K., 1944. *Trans. Amer. Inst. Elect. Engrs.*, **63**, 251.
HOFFERT, H. H., 1890. *Proc. Phys. Soc.*, **10**, 176.
HOLZER, R. E. and WORKMAN, E. J., 1939. *J. Appl. Phys.*, **10**, 659.
HOLZER, R. E., WORKMAN, E. J. and SNODDY, L. B., 1938. *J. Appl. Phys.*, **9**, 134.
HUMPHREYS, W. J., 1936. *Proc. Amer. Phil. Soc.*, **76**, 613.
KUEHNI, H. P. and RAMO, S., 1937. *Trans. Amer. Inst. Elect. Engrs.*, **56**, 721.
LABY, T. H., MCNEILL, J. J., NICHOLLS, F. G. and NICKSON, A. F. B, 1940. *Proc. Roy. Soc.*, A, **174**, 145.
LARSEN, A., 1905. *Annual Report Smithsonian Institute*, p. 119.
LEE, E. S., 1928. *Gen. Elect. Rev.*, **31**, 404.
LEE, E. S. and FOUST, C. M., 1927. *Trans. Amer. Inst. Elect. Engrs.*, **46**, 149.
LENARD, P., 1892. *Wied. Ann.*, **46**, 584.
LEWIS, W. W. and FOUST, C. M., 1930. *J. Amer. Inst. Elect. Engrs.*, **49**, 389.
LEWIS, W. W. and FOUST, C. M., 1945. *Trans. Amer. Inst. Elect. Engrs.*, **64**, 107.
LOEB, L. B. and KIP, A. F., 1939. *J. Appl. Phys.*, **10**, 142.
LOEB, L. B. and MEEK, J. M., 1940. *J. Appl. Phys.*, **11**, 435 and 459.
LUTKIN, F. E., 1937. *J. Sci. Instrum.*, **14**, 209.
LUTKIN, F. E., 1938. *J. Instn. Elect. Engrs.*, **82**, 189.
LUTKIN, F. E., 1939. *Proc. Roy. Soc.*, A, **171**, 285.
LUTKIN, F. E., 1941. *Quart. J. R. Met. Soc.*, **67**, 347.
MALAN, D. J. and COLLENS, H., 1937. *Proc. Roy. Soc.*, A, **162**, 175.
MATTHIAS, A., 1929. *Elektrotech. Z.*, **50**, 1469.
MATTHIAS, A., 1937. *Elektrotech. Z.*, **58**, 881, 928 and 973.
MATTHIAS, A. and BURKHARDTSMAIER, W., 1939. *Elektrotech. Z.*, **60**, 681 and 730.
MCCANN, G. D., 1942 a. *J. Instn. Elect. Engrs.*, **89**, 646.

McCann, G. D., 1942 b. Paper presented to Inst. of Aeronautical Sciences (America), Dec. 15th, 1942.
McCann, G. D., 1944. *Trans. Amer. Inst. Elect. Engrs.*, **63**, 1157.
McCann, G. D. and Clark, J. J., 1943. *Trans. Amer. Inst. Elect. Engrs.*, **62**, 45.
McCann, G. D. and Morgan, D. E., 1943. *Trans. Amer. Inst. Elect. Engrs.*, **62**, 427.
McEachron, K. B., 1939. *J. Franklin Inst.*, **227**, 149.
McEachron, K. B., 1941. *Trans. Amer. Inst. Elect. Engrs.*, **60**, 885.
McEachron, K. B. and Hagenguth, J. H., 1942. *Trans. Amer. Inst. Elect. Engrs.*, **61**, 559.
McEachron, K. B. and Thomason, J. L., 1935. *Gen. Elect. Rev.*, **38**, 126.
McMorris, W. A., Rusher, M. A. and Hagenguth, J. H., 1934. *Gen. Elect. Rev.*, **37**, 514.
Meek, J. M., 1941. *Nature, Lond.*, **148**, 437.
Meek, J. M., 1942. *J. Instn. Elect. Engrs.*, **89**, 335.
Meek, J. M. and Craggs, J. D., 1943. *Nature, Lond.*, **152**, 538.
Merrill, F. H. and von Hippel, A., 1939. *J. Appl. Phys.*, **10**, 873.
Miranda, F. J., 1937. *J. Instn. Elect. Engrs.*, **81**, 46.
Neugebauer, T., 1937. *Z. Phys.*, **106**, 474.
Norinder, H., 1928. *Trans. Amer. Inst. Elect. Engrs.*, **47**, 446.
Norinder, H., 1935. *J. Franklin Inst.*, **220**, 63.
Norinder, H., 1937. *Proc. Phys. Soc.*, **49**, 364.
Noto, H. 1934. *Proc. Phys.-Math. Soc. Japan*, **16**, 177.
Perry, F. R., 1941. *J. Instn. Elect. Engrs.*, **88** (Part II), 69.
Perry, F. R., 1944. *Quart. J. R. Met. Soc.*, **70**, 169.
Perry, F. R. and Jobling, H. B., 1936. *Trans. S. Afr. Inst. Elect. Engrs.*, **27**, 41.
Perry, F. R., Webster, G. H. and Baguley, P. W., 1942. *J. Instn. Elect. Engrs.*, **89** (Part II), 185.
Peters, J. F., 1924. *Elect. World, N.Y.*, **83**, 769.
Pittman, R. R. and Torok, J. J., 1931. *Trans. Amer. Inst. Elect. Engrs.*, **50**, 568.
Schonland, B. F. J., 1928 a. *Proc. Roy. Soc.*, A, **118**, 233.
Schonland, B. F. J., 1928 b. *Proc. Roy. Soc.*, A, **118**, 252.
Schonland, B. F. J., 1937. *Phil. Mag.*, **23**, 503.
Schonland, B. F. J., 1938 a. *Proc. Roy. Soc.*, A, **164**, 132.
Schonland, B. F. J., 1938 b. *The Lightning Discharge* (Oxford: University Press).
Schonland, B. F. J., 1938 c. *Nature, Lond.*, **141**, 115.
Schonland, B. F. J. and Allibone, T. E., 1931. *Nature, Lond.*, **128**, 794.
Schonland, B. F. J. and Collens, H., 1934. *Proc. Roy. Soc.*, A, **143**, 654.
Schonland, B. F. J. and Craib, J., 1927. *Proc. Roy. Soc.*, A, **114**, 229.
Schonland, B. F. J., Elder, J. S., Hodges, D. B., Phillips, W. E. and van Wyk, J. W., 1940. *Proc. Roy. Soc.*, A, **176**, 180.
Schonland, B. F. J., Elder, J. S., van Wyk, J. W. and Cruickshank, G., 1939. *Nature, Lond.*, **143**, 893.
Schonland, B. F. J. and Hodges, D. B., 1936. *Trans. Roy. Soc. S. Afr.*, **24**, 81.
Schonland, B. F. J., Hodges, D. B. and Collens, H., 1938. *Proc. Roy. Soc.*, A, **166**, 56.
Schonland, B. F. J., Malan, D. J. and Collens, H., 1935. *Proc. Roy. Soc.*, A, **152**, 595.
Schonland, B. F. J., Malan, D. J. and Collens, H., 1938. *Proc. Roy. Soc.*, A, **168**, 455.
Shipley, J. F., 1940. *Distrib. Elect.*, **12**, 413.
Simpson, G. C., 1909. *Phil. Trans.*, A, **209**, 412.
Simpson, G. C., 1919. *British Antarctic Expedition* 1910-13: *Meteorology, Calcutta*, **1**, 306.
Simpson, G. C., 1927. *Proc. Roy. Soc.*, A, **114**, 376.
Simpson, G. C., 1942. *Quart. J. R. Met. Soc.*, **68**, 1.
Simpson, G. C. and Robinson, G. D., 1941. *Proc. Roy. Soc.*, A, **177**, 281.
Simpson, G. C. and Scrase, F. J., 1937. *Proc. Roy. Soc.*, A, **161**, 309.
Smeloff, N. N. and Price, A. L., 1930. *Trans. Amer. Inst. Elect. Engrs.*, **49**, 895.
Sporn, P. and Lloyd, W. L., 1930. *Trans. Amer. Inst. Elect. Engrs.*, **49**, 905.
Stekolnikov, I. S., 1937 a. *Conf. int. gr. Rés. Élect.*, Paper No. 129.
Stekolnikov, I. S., 1937 b. *Conf. int. gr. Rés. Élect.*, Paper No. 327.
Stekolnikov, I. S. and Beliakov, A. P., 1936. *Elektrichestvo*, **57** (section 22), 16.
Stekolnikov, I. S. and Lamdon, A. A., 1938. *Elektrichestvo*, **59** (section 5), 29.

STEKOLNIKOV, I. S. and VALEEV, C., 1937. *Conf. int. gr. Rés. Élect.*, Paper No. 330.
STEKOLNIKOV, I. S. and VALEEV, C., 1938. *Elektrichestvo*, **59** (section 1), 11.
STEKOLNIKOV, I. S. and YAVORSKY, V. V., 1936. *Elektrichestvo*, **57** (section 8), 13.
THORNTON, W. M., 1911. *Phil. Mag.*, **21**, 630.
TOROK, J. J. and FIELDER, F. D., 1929. *Elect. J.*, **26**, 320.
WAGNER, C. F. and MCCANN, G. D., 1940. *Trans. Amer. Inst. Elect. Engrs.*, **59**, 1061.
WAGNER, C. F. and MCCANN, G. D., 1941. *Elect. Engng., N.Y.*, **60**, 374, 438 and 483.
WAGNER, C. F., MCCANN, G. D. and BECK, E., 1941. *Trans. Amer. Inst. Elect. Engrs.*, **60**, 1222.
WAGNER, C. F., MCCANN, G. D. and LEAR, C. M., 1942. *Trans. Amer. Inst. Elect. Engrs.*, **61**, 96.
WAGNER, C. F., MCCANN, G. D. and MACLANE, G. L., 1941. *Trans. Amer. Inst. Elect. Engrs.*, **60**, 313.
WALTER, B., 1903. *Ann. Phys., Lpz.*, **10**, 393.
WALTER, B., 1935. *Ann. Phys., Lpz.*, **22**, 421.
WALTER, B., 1936. *Ann. Phys., Lpz.*, **25**, 124.
WATSON WATT, R. A., HERD, J. F. and BAINBRIDGE-BELL, L. H., 1933. *Applications of the Cathode-ray Oscillograph in Radio Research* (London: H.M. Stationery Office).
WATSON WATT, R. A., HERD, J. F. and LUTKIN, F. E., 1937. *Proc. Roy. Soc.*, A, **162**, 267.
WHIPPLE, F. J. W. and SCRASE, F. J., 1936. *Meteor. Off. Geophys., Mem.* 7, No. 68.
WILSON, C. T. R., 1916. *Proc. Roy. Soc.*, A, **92**, 555.
WILSON, C. T. R., 1921. *Phil. Trans.*, A, **221**, 73.
WILSON, C. T. R., 1929. *J. Franklin Inst.*, **208**, 1.
WOLF, F., 1943. *Naturwissenschaften*, **31**, 215.
WORKMAN, E. J., BEAMS, J. W. and SNODDY, L. B., 1936. *Physics*, **7**, 375.
WORKMAN, E. J. and HOLZER, R. E., 1939 a. *Rev. Sci. Instrum.*, **10**, 160.
WORKMAN, E. J. and HOLZER, R. E., 1939 b. *Int. Geophys. Un.*, September.
WORMELL, T. W., 1927. *Proc. Roy. Soc.*, A, **115**, 443.
WORMELL, T. W., 1930. *Proc. Roy. Soc.*, A, **127**, 567.
WORMELL, T. W., 1939. *Phil. Trans.*, A, **238**, 249.
ZADUK, H., 1935. *Arch. tech. Messen*, **5**, 155.
ZALESSKI, A. M., 1935. *Conf. int. gr. Rés. Élect.*, Paper No. 317.

SUPERCONDUCTIVITY

By K. MENDELSSOHN,
Clarendon Laboratory, Oxford

§1. INTRODUCTION

THE last of these *Reports* on superconductivity (Jackson, 1939) dealt with material published up to the end of 1939. The position in the intervening war years, which are covered by the present *Report*, has been somewhat anomalous. Research on superconductivity was completely stopped in this country, in which, previously, a considerable proportion of the work had been carried out, while it continued at even more than the peace-time rate in Germany. The Leyden laboratory has kept up research in the early years of occupation, but later had to close down almost completely. In spite of the war, a great deal of work has been carried out in the Soviet Union. No papers on superconductivity have been published from the Toronto laboratory. Owing evidently to inadequate communication, many authors appear to be unaware of contemporary and earlier work, and there is a certain amount of duplication of earlier researches. In an effort to present a continuous and balanced picture of the present status of knowledge of superconductivity, occasional reference has been made in this *Report* to work published before 1940, so as to link up and compare the new results with those of earlier years. The services of scientific abstracts having broken down to some extent, considerable difficulty was experienced in collecting the material, and no claim can be made for completeness of the list of references.

§2. CRITICAL DATA

The superconductive properties of a number of metals have been investigated by Shoenberg (1940a). The threshold curves of aluminium, gallium and thorium were established between T_c* and 0·98°K., and that of zinc between T_c and 0·55°K. The latter results were obtained by using the magnetic method of cooling, and are subject to a larger error than those taken in liquid helium. A graphic representation of the results is given in figure 1. The behaviour of two different samples of thorium is of particular interest. While one showed a magnetization curve similar to that of alloys, another, purer one, behaved like other very pure metals. This is exactly the same behaviour as had been observed in tantalum (Mendelssohn and Moore, 1936 ; Daunt and Mendelssohn, 1937). A sample of titanium did not show signs of superconductivity except for a slight magnetic effect which appeared below 1·5°K., although this metal is listed as a superconductor (Meissner, 1930). Similar indications of superconductivity were found in a sample of uranium at 1·1°K. Molybdenum is not superconductive at temperatures above 0·3°K.

Superconductivity of rhenium at 0·95°K. has been reported by Aschermann

* We use the following notation throughout this *Report* : T_c for the transition temperature in zero magnetic field, H_c for the critical field (threshold value) and I_c for the critical current.

and Justi (1942), who also mention in passing the superconductivity of uranium at 1·3° K. without giving further details. The superconductivity of uranium, and even more that of rhenium, is certainly surprising, since these metals stand

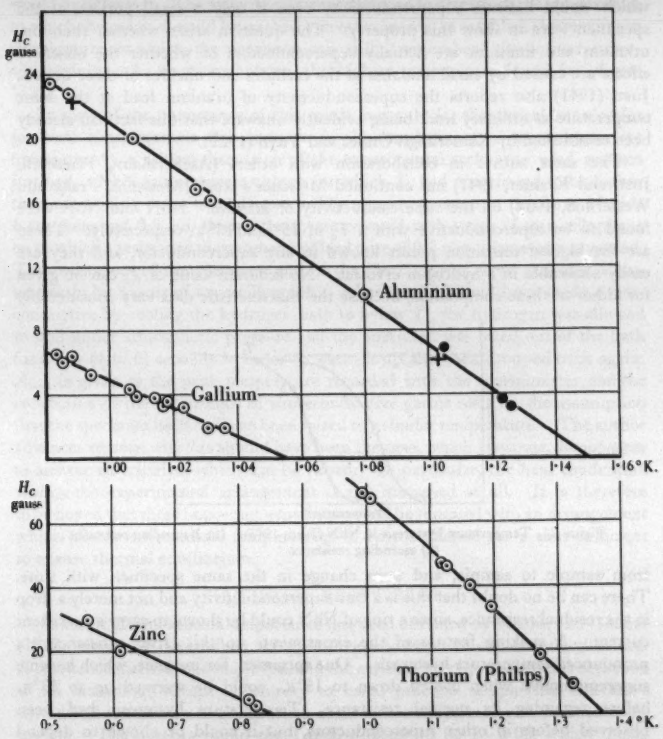


Figure 1. Threshold curves of aluminium, gallium, thorium and zinc (Shoenberg, 1940 a).

(*Reproduced, by kind permission, from the 'Proceedings of the Cambridge Philosophical Society'.*)

in groups VI and VII of the periodic table, which contain only non-superconductive metals.* The authors draw the conclusion that at sufficiently low temperatures all metals may become superconductive, and not only a certain

* It should, however, be noted that the recently determined chemical properties of the nsuranic elements seem to indicate that uranium may be a member of a new rare earth group which starts with actinium. This means that the electronic configuration of uranium need not necessarily correspond to group VI of the periodic table.

section covered by groups II *b* to V *a*. It must, however, be noted that while Shoenberg has measured magnetic susceptibilities, Justi has found the superconductivity of rhenium in a determination of the electrical resistance, a method which would indicate superconductivity even if only a small portion of the specimen were to show this property. The question arises whether rhenium, uranium and titanium are actually superconductors or whether the observed effects are caused by small amounts of the carbides and nitrides of these metals. Justi (1941) also reports the superconductivity of uranium lead at the same temperature as ordinary lead, being evidently unaware that this fact had already been established by Kamerlingh Onnes and Tuyn (1922).*

The same author in collaboration with others (Aschermann, Friedrich, Justi and Kramer, 1941) has continued Meissner's work (Meissner, Franz and Westerhoff, 1932) on the superconductivity of nitrides. NbN and NbH were found to be superconductive with a T_c of 15° and 13° K. respectively. These are the highest transition points known in any superconductor, and they are easily accessible in a hydrogen cryostat. No accurate value of T_c can be given for either of these compounds, because the characteristic data vary considerably

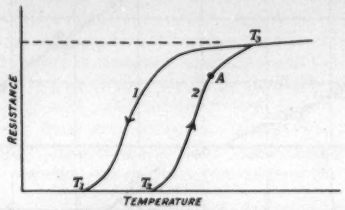


Figure 2. Temperature hysteresis in NbN (Justi, 1943). (1) descending resistance; (2) ascending resistance.

from sample to sample, and even change in the same specimen with time. There can be no doubt that this is a true superconductivity and not merely a drop in the residual resistance, since a ring of NbN could be shown to carry a persistent current. A striking feature of the experiments on this latter substance is a pronounced temperature hysteresis. One specimen, for instance, which became superconductive when cooled down to 15° K., could be warmed up to 23° K. before regaining its normal resistance. Temperature hysteresis had been observed before in other superconductors, but it could be shown to depend on the strength of the measuring current, and disappeared when the current was reversed during the experiment. No such effect was observed in NbN. From a second series of experiments on NbN Justi (1943) concludes that superconductive "germs" are retained in the specimen up to the temperature of liquid air and even higher (112° K.). The argument is as follows. A specimen of NbN will lose its resistance completely at T_1 (figure 2), which corresponds to a temperature of about 15° K. On warming, the specimen as a whole will stop being superconductive at T_2, but the normal resistance will only reappear in full at T_3.

* Superconductivity of the compound Bi_3Ni, consisting of two non-superconductive components, has recently been reported by Alekseyevsky (1945 c).

If the specimen is warmed up to the point A above T_2 and subsequently cooled, the resistance will follow the ascending curve (2) as long as superconductive "germs" are present, while after the complete destruction of these germs curve (1) would be followed. The experiments were designed to determine the temperature to which the point A could be pushed ; in other words, to establish the maximum temperature at which the specimen can still "remember" its former superconductivity. As mentioned above, the highest temperature found for A was 112° K.

In view of the importance of this result, not only for theoretical but also for practical considerations (for instance in the employment of superconductive bolometers *), a critical discussion of the experimental technique used appears justified. NbN samples were chosen for which T_2 had been found to be higher than the boiling point of hydrogen. The specimens were in the form of rods 6 cm. long and 0·3 cm. thick ; they were covered with strips of thin cellophane on to which a resistance thermometer of lead wire of 0·2 mm. diameter was wound. The resistances of the specimen and of the thermometer were recorded simultaneously by means of an oscillograph. After the NbN had been made superconductive by cooling the hydrogen bath to below T_1, the hydrogen was allowed to boil under atmospheric pressure and the specimen was lifted out of the bath for a few (4 to 6) seconds in order to warm it up and then dropped back again. A_{max} is given as the peak temperature recorded with the thermometer, and the conclusion of the persistence of superconductive germs rests on the assumption that the specimen itself had also been raised to a similar temperature. The author advances reasons why this should have been the case, which, however, do not seem to answer all criticism which can be raised. In particular, the heat conduction within the experimental arrangement is not discussed at all. It is therefore to be hoped that these important experiments will be repeated with an arrangement which allows intermediate temperatures to be maintained for a time sufficient to ensure thermal equilibrium.

§ 3. BASIC PROPERTIES

A result of fundamental importance has been obtained by Kikoin and Goobar (1940), who have determined the gyro-magnetic effect in superconductive lead. A resonance method was employed, and the experiment required careful exclusion of all disturbing effects. Contrary to usual practice, a spherical and not a cylindrical specimen was used, which ensured better symmetry of the measuring system. The authors found the Landé g-factor to be $1 \pm 0{\cdot}03$, which means that the diamagnetism of superconductors is produced by ordinary electron currents and not, as has sometimes been suggested, by electron spin. While the result has thus led to a definite answer as to the nature of the diamagnetic current, it has also introduced a new difficulty. The phenomenon of zero resistance, and in particular the absence of a Thomson heat (Daunt and Mendelssohn, 1938), seem to be clear indications that no interaction takes place between the electrons engaged in super-currents and the lattice. Under these circumstances it is hard to understand why a superconductor should exhibit

* Brucksch and Ziegler, 1942 ; Ziegler, Brucksch and Hickman, 1942.

the gyro-magnetic effect at all. One possible explanation is the existence of "frozen-in" lines of force in the specimen, in spite of the fact that the authors used very pure lead and compensated the earth's magnetic field. Another possibility, not mentioned by the authors, is that in polycrystalline material an additional energy may be needed to shift a surface current over a crystal boundary. Indications of the existence of such an effect were observed in experiments at Oxford, which have not been published.

According to Silsbee's hypothesis (Silsbee, 1916), the limiting current I_c is that which produces the magnetic field H_c on the surface of the superconductor. A test of this hypothesis was made by Tuyn and Kamerlingh Onnes (1926), and has now been repeated by Steiner and Gerschlauer (1942). If a current is passed through a wire running along the axis of a superconductive tube, the resistance of the tube should show a minimum if the field produced by the wire compensates the field caused on the surface of the tube by its own measuring current. The results gave, indeed, such a minimum in all but one of the cases (tin cylinders) investigated. The numerical agreement with the hypothesis is, however, not so satisfactory, since these minima occur at compensating fields appreciably smaller than calculated. No explanation for the discrepancy is offered, but one may suspect that it is due to the somewhat intricate problem of current distribution in the actual transition region.

§4. ANOMALOUS SUPERCONDUCTORS

As in previous years, the great bulk of the experimental work on superconductors is concerned with the mechanism of the transition between the superconductive and the normal phase and its dependence upon structure. Nakhutin (1942) has measured the magnetic moment of a tin single crystal in different directions, and has found that H_c is independent of the direction of the crystal axis. This is in good accord with the earlier experiments of de Haas and Voogd (1931), who found that the change in electrical resistance is uninfluenced by the crystal axis. The continuity of the transition had already been investigated by Shubnikov, Chotkevitch, Shepelev and Rjabinin (1936), who had found no jumps, similar to those observable in ferromagnetics, in the magnetization of superconductors. It was therefore surprising when Justi (1942 a) announced the discovery of such discontinuities in NbN and pure niobium and, in a later communication (1942 b), in a cylinder and a sphere of tin. In the first experiments the jumps were registered as noise and in the later ones oscillographically recorded. The discrepancy between these results and the earlier observations appears to be explained by the work of Alekseyevsky (1945 b), who has carried out experiments similar to Justi's on polycrystalline and monocrystalline cylinders. He found that discontinuities only occur when polycrystalline material is used or in monocrystalline specimens if they are investigated in a slightly inhomogeneous magnetic field; unfortunately this author gives no indication as to the metal investigated or the ranges of temperature and magnetic field employed. Since Shubnikov had taken particular care in his experiments to ensure homogeneity of the field, and since Justi employed polycrystalline specimens, worked in a lathe, it seems safe to conclude that the

discontinuities observed by the latter are not a fundamental feature of the transition mechanism but are due to the influence of inhomogeneities. This conclusion is corroborated by experiments of Steiner and Schoeneck (1943 a), who also observed jumps in polycrystalline samples but a smooth transition in single crystals. The main object of their experiments, done on tin and indium, was the simultaneous observation of the changes in resistance and magnetic induction occurring at the transition from the normal to the superconductive state. The transition was effected both by changing the temperature at constant field and by decreasing the field at constant temperature. The results show quite generally that the first drop in the resistance is accompanied by only a small decrease in magnetic induction, and that the magnetic field is completely pushed out of the specimen (Meissner effect *) well before the resistance has attained the value zero. The authors do not comment on the outcome of their work, which, however, must seem to some extent incompatible with the conceptions held so far. Since the measurements were carried out in a longitudinal field, one should have expected the formation of superconductive needles in the direction of the measuring current, causing zero resistance, even before the change in induction is completed. A possible explanation for the observed phenomena is the influence of the field produced by the measuring current. This field was, in the arrangement of Steiner and Schoeneck, perpendicular to the external field, and changes in its distribution would not be recorded in their search coil. The same influence may be the cause of a still more peculiar effect mentioned by the same authors in a subsequent paper (Steiner and Schoeneck, 1943 b). Employing the same experimental arrangement, but using a stronger measuring current, they observed at the onset of the transition a slight *increase* in the induction of the specimen which preceded the change to zero induction. Unless one is prepared to assume the existence of an entirely new and, so far, unknown effect in the mechanism of the transition from the normal to the superconductive state, it must be suspected that no such paramagnetic change does actually take place, but that the increase in induction is only apparent, and is caused by the redistribution of the measuring currrent and its attendant field. It is clear that more work is required before this point can be settled.

Two determinations of the specific heat of a superconductor have been carried out, both on tantalum. In pure tin the difference between the specific heats of the metal in the superconductive and the normal phase (Keesom and van Laer, 1938) can be expressed in terms of H_c (Gorter and Casimir, 1933), whereas no such relation holds in alloys (Mendelssohn and Moore, 1935). Fairly pure tantalum specimens had shown a magnetic and electric behaviour indistinguishable from that of alloys (Mendelssohn and Moore, 1936), while a specimen of very high purity had yielded results similar to those found in other pure metals (Daunt and Mendelssohn, 1937). The anomalous behaviour of alloys and some tantalum specimens (see also the results on thorium (Shoenberg, 1940 a) mentioned above) has been ascribed to the existence of a superconductive "sponge" (Mendelssohn, 1935) consisting of small regions of high threshold value, which is formed by the internal strain due to chemical or physical

* Meissner and Ochsenfeld, 1933.

impurities. Keesom and Désirant (1941) have measured the specific heat of a tantalum specimen, whose magnetic properties were similar to those of alloys, in the normal and in the superconductive state (figure 3). The absence of

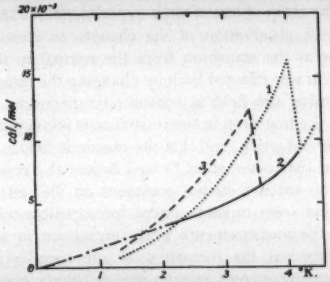


Figure 3. Specific heats of tantalum (Keesom and Désirant, 1941). (1) in zero magnetic field, (2) in a field of 3000 gauss, and (3) in a field of 691·6 gauss.

a latent heat at the passage from the superconductive to the normal state in a magnetic field (curve 3) is only in part due to the geometric shape of the specimen. It is a feature of the sponge structure that the field penetrates the specimen

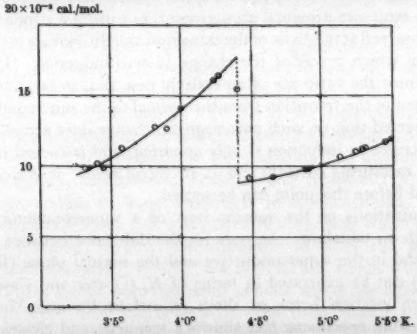


Figure 4. Specific heat of tantalum in zero magnetic field as measured by Désirant and Mendelssohn (Mendelssohn, 1941).
(*Reproduced, by kind permission, from 'Nature'.*)

gradually, and it is evidently owing to this gradual transition that no distinct onset of a latent heat was observed. In zero field the transition is spread over 0·2°, being completed at 4·155° K.* Désirant and Mendelssohn (Mendelssohn, 1934) have measured the specific heat of the very pure sample, which in its magnetic behaviour resembles other pure superconductors. The determinations

* Impurities and stain may shift T_c to higher or lower temperatures

were made in zero field and in the neighbourhood of the transition point (figure 4). In this sample the transition region was found to be much narrower (of the order of a few hundredths of a degree) and to lie at 4·4° K., which is in good agreement with the resistance change in a very pure wire (Meissner, 1928). The discontinuity at T_c amounts to 9×10^{-3} cal./mol, as predicted from the magnetic measurements of Daunt and Mendelssohn (1937). The two sets of experiments show clearly that a sufficiently pure and homogeneous specimen of tantalum behaves similarly to other pure superconductors, and that the anomalous behaviour of some samples is due to secondary causes (strain).

Direct evidence that the sponge structure of certain superconductive specimens is caused by strain has been provided by an ingenious experiment carried out by Lasarew and Galkin (1944). They investigated the resistance of tin wires (~0·05 mm. diameter) which were glued with celluloid on to glass plates. Since the thermal expansion of glass is considerably smaller and that of celluloid much larger than that of tin, the wires were stretched in the direction of the axis and compressed in the direction of the radius when the specimens were cooled. At the same time specimens of the unstrained wire were examined as controls. The ratio of resistance at 4·2° K. to that at room temperature was determined as well as T_c, H_c and I_c. The results are given in table 1.

Table 1

Specimen	$R_{4\cdot2}/R_{\text{room}}$	T_c °K.	H_c (2° K.) gauss	dH_c/dT (2° K.)	I_c (2° K.) amperes
Unstrained	$1\cdot5 \times 10^{-2}$	3·72	210	100	3·000
Strained	0·1	9·00	15000	1750	0·067

The most striking effect of strain is the enormous increase in T_c, showing that the transition temperature has been more than doubled. The drop in I_c and the rise in H_c are equally impressive. In fact one can say that all the characteristic properties of a superconductive alloy have been reproduced in a chemically pure specimen by the application of strain. In addition, considerable hysteresis effects at the disturbance of superconductivity by a current were observed. The authors point out that the results cannot be simply explained as being due to the extension of the wire. Using the known values (Sizoo and Kamerlingh Onnes, 1925 ; Lasarew and Kan, 1944 b) for the shift of T_c with deformation, they calculate that a pressure of 90,000 kg./cm² would have been necessary to produce the observed shift of 5°. On the other hand, the same pressure would only account for a change in H_c of 700 gauss and not, as observed, of 15,000 gauss. They conclude that, apart from the extension of the metal lattice, there must be in operation another process, which they believe to be the splitting up of the superconductor into small domains. This is, of course, what is to be expected according to the sponge hypothesis. It is interesting to note in this connection that the specimens of NbN which show a high T_c are also distinguished by high H_c and low I_c (Aschermann, Friedrich, Justi and Kramer, 1941).

Lasarew and Kan have employed another ingenious method (1944 a) to examine (1944 b) the effect of an all-sided pressure of 1750 kg./cm^2 on the magnetic induction and the resistance of tin and indium when passing into the superconductive state. The specimens were enclosed at room temperature in a pressure cylinder filled with water, and the rise of pressure on freezing was utilized. The transition temperature of tin in zero magnetic field was found to be lowered by $0{\cdot}095 \pm 0{\cdot}005°$. The authors conclude that up to 2000 kg./cm^2 T_c and H_c vary in proportion to the pressure applied.

The change of T_c under stress has also been measured by Alekseyevsky (1945 a), who has determined simultaneously the variation of resistance and induction of tin single crystals subjected to a stress of 0·64 kg./mm^2 and has found them to be shifted by 0·001°. His value of dT_c/dp is slightly lower than the one observed by Sizoo and Kamerlingh Onnes (1925) for polycrystalline tin. The same author has also measured the displacement of the critical data of tantalum under tension (1940).

In previous years the behaviour of very thin films of mercury had been thoroughly investigated by Appleyard, Bristow, London, H., and Misener (1939). This work has now been repeated with lead films by Brucksch and Ziegler (1942). Their results are, on the whole, in excellent agreement with those on mercury, showing a decrease in I_c and an increase in H_c with decreasing film thickness, while T_c remains the same as in the bulk metal. However, the H_c/T-curves are concave upwards, a feature not observed with the mercury films and very uncommon in superconductors. The only other case known is that of a tantalum wire examined by Mendelssohn and Moore (1936). It is interesting that this particular specimen showed alloy behaviour. Ziegler, Brucksch and Hickman (1942) have given a number of useful data on wires of Sn, Nb, Ta and Pb varying in thickness between 25 and 250 μ. The object of this work, as well as of that on lead films, was to investigate the suitability of various specimens for use in a superconductive bolometer.

The only work on superconductive alloys reported is a series of magnetic experiments on tin-zinc alloys carried out by Lasarew and Nakhutin (1942). The induction of three disk-shaped specimens with a tin content of 90 %, 34 % and 10 % was determined in a magnetic field parallel to the face of the disks. The case under investigation differs from previous experiments in that the two constituents do not form a solid solution. All specimens showed signs of superconductivity up to the same fields as pure tin, but, as the zinc content is increased, lines of force begin to penetrate the specimens at progressively lower fields. There is also a marked hysteresis which increases with the zinc content. All these phenomena can be explained by assuming the existence of multiply-connected superconductive-regions of pure tin which are embedded in the non-superconductive zinc. In this way the tin-zinc alloys exhibit "freezing in" of magnetic flux, as is to be expected of a sponge structure, without, however, deviating from the H_c/T-curve of pure tin, because the superconductive regions are evidently not of very small size. The results are valuable because they furnish an intermediate stage between the behaviour of pure metals and that of alloys.

Summarizing the outcome of the work on the specific heat of tantalum, the effect of strain, the behaviour of thin films and of the tin-zinc alloys, it can be said that all these results present a fairly consistent picture of the general trend of deviations from the behaviour of "ideal" superconductors, as, for instance, a specimen of pure tin, which has a simple geometrical shape. It seems that the sponge hypothesis, postulating the existence of multiply-connected superconductive-regions of small size which are caused by strain, is well borne out by the experiments. However, the question *why* strain should cause the magnetically homogeneous structure of the ideal superconductor to split up remains open. All available evidence seems to point to the existence of a "surface tension" of the superconductive against the normal phase.

§ 5. THE INTERMEDIATE STATE

This surface tension plays an important part in the interpretation of the intermediate state, i.e. that region of the H/T diagram of an ideal superconductor for which the gross magnetic induction has a value intermediate between zero and H_c. Landau, following his earlier papers (1937, 38), has published (1943) a detailed theory of the intermediate state based on thermodynamical considerations. According to this, a superconductor in the intermediate state is supposed to consist of normal and superconductive laminae, the relative thicknesses of which depend on the induction of the specimen, that is, on the external magnetic field. The total thickness of a layer consisting of a normal and a superconductive lamina is field independent, and is of the order

$$\sim 2d^{\frac{1}{3}}L^{\frac{1}{3}}$$

in a plate of the thickness L which stands normal to the direction of the field. The quantity d has the dimension of a length and depends on the surface tension between the normal and the superconductive phase and the inverse square of the threshold field ; it is of the same order of magnitude as the depth to which a magnetic field can penetrate a superconductor (Becker, Heller and Sauter, 1933 ; London, H., 1934). Assuming for d the value of $\sim 3 \times 10^{-5}$ cm., the layers will be of the order of 0·05 cm. in a plate of 1 cm. thickness. The laminae are not stable in the neighbourhood of the metal surface, but must branch successively into thinner and thinner laminae. In a plate of 1 cm. thickness the branching should begin at a depth of 0·14 cm. below the surface. While branching must prevent any observation of the laminar structure on the free surface of a superconductor, its existence may be established by observations in narrow slits in the metal which stand normal to the direction of the lines of force, i.e. perpendicular to the laminae. In a superconductive plate of 1 cm. thickness the permitted width of the slit is estimated to be 0·1 mm. According to Landau, the laminar structure of the intermediate state should also cause a persistence of superconductivity to fields higher than those usually calculated from the geometrical data of the specimen. For instance, in a cylinder of circular cross-section, whose axis is perpendicular to the direction of the magnetizing field, the onset of the intermediate state should occur at $\frac{1}{2}H_c$ if the external field is measured at infinite distance from the specimen. At this stage the field on the surface of the specimen

will have reached the value H_c. Landau postulates that the change should only take place at the higher field,

$$H_c\left(\frac{1}{2}+\cdot 4\sqrt{\frac{d}{r}}\right),$$

where r is the radius of the cylinder. For a sphere with the radius r the field of first penetration is calculated as

$$H_c\left(\frac{2}{3}+\cdot 5\sqrt{\frac{d}{r}}\right).$$

Since d is defined as being proportional to $1/H_c^3$, the effect should be most marked in the vicinity of T_c.

Experiments to observe the laminar structure, and to test Landau's theory, have been undertaken by Shalnikov (1945), who has measured the induction in narrow slits cut along the equatorial plane of superconductive tin spheres. The field inside the spheres was examined by measuring the resistance of bismuth

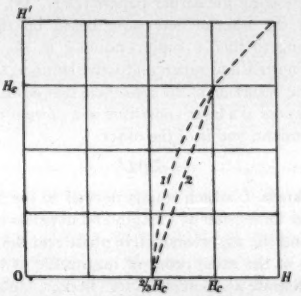


Figure 5. Magnetic field in the equatorial plane of a superconductive sphere (Shalnikov, 1945). Width of slit (1) 20 μ and (2) 100 μ.

probes of 0·3 to 0·5 mm. in width and 4 to 6 μ in thickness. The position of the probe was fixed, but the width of the slit could be varied, the minimum width being about 20 μ. Figure 5 shows the change of the internal field, as measured by the probe, with increasing external field. In accordance with the results obtained by previous observers (Mendelssohn and Babbitt, 1935; Shoenberg, 1935; de Haas and Guinau, 1936), the first penetration occurs at about $\frac{2}{3}H_c$ and the internal field is equal to the external one above H_c. While however, also in accordance with earlier work, it was found that the internal field rises linearly with the external field in a slit 100 μ in width, this rise is curved if the diameter of the slit is reduced to 20 μ. Shalnikov points out that, while it would be impossible, with the experimental arrangement employed, to demonstrate the existence of individual laminae, the curved rise of the internal field is exactly what would have to be expected in a slit narrow enough for the laminar structure to be maintained.

Increasing the width of the slit to 100 μ on the other hand destroys the laminar structure on the free surface of the metal. From a number of experiments of this kind the author estimates the critical width as being between 30 to 50 μ, from which the surface tension between the normal and the superconductive phase is calculated as $1{\cdot}7 \times 10^{-3}$ to $4{\cdot}4 \times 10^{-3}$ dyne/cm. While no direct observation of the individual layers was possible, their size could be calculated by using the deduced value of the surface tension. The thickness of these layers, consisting of one superconductive and one normal lamina, comes to 0·4 to 0·63 mm.

It is interesting to compare these results with those of de Haas and Guinau (1936), who passed a probe through a channel along the equatorial plane of a monocrystalline sphere. These authors did actually observe fluctuations of the internal field which seem to correspond to alternating superconductive and normal laminae, whose thickness was, however, much larger (of the order of 1 cm.).

Shalnikov (1942) has also investigated the mechanism of the magnetic transition of hollow tin spheres. The method employed and the results for the equatorial field are practically identical with those of Mendelssohn and Babbitt (1935). The present experiments are extended to field measurements at different latitudes on spheres of varying wall-thickness. It may be expected that at the destruction of the superconductive state in a hollow sphere, an annular zone covering the equator will retain superconductivity up to fairly high external fields, while at higher latitudes superconductivity is destroyed. This effect could actually be observed, but it is pointed out that it is not possible to account fully for all the observed phenomena. This is hardly surprising in view of the complicated geometrical problem presented by the magnetic transition in a hollow sphere. Experiments on the intermediate state of a superconductive ellipsoid have been reported by Nakhutin (1939).

The question of the onset of the intermediate state at tangential fields higher than H_c has also been examined theoretically by Casimir (1943), who has derived the free energy of the transition between the superconductive and the normal state of a sphere. His calculations show that the homogeneous solution of the intermediate state yields an absolute minimum of the free energy, leading nowhere to the appearance of a surface current. On the other hand, the experimental evidence strongly suggests the existence of such currents. It is thought that these currents may occur in the mixed phase which is identical with the region, near the surface, in which the branching of Landau's laminae takes place. An important conclusion reached is that the surface currents, which must introduce irreversible effects, are responsible for the time effects observed in superconductors. This explanation had been suggested, and had actually been demonstrated in specimens of irregular shape (Mendelssohn and Pontius, 1936 a, 36 b). However, it has been pointed out (van Laer and Groenewold, 1939) that the time effects in an ellipsoid, presenting a simple geometrical case, could hardly be attributed to the dying out of persistent currents. Casimir's conclusions indicate that, nevertheless, this seems to be the case. By thus attributing all the time effects observed in different investigations to one and the same cause, the problem of these puzzling effects is much simplified.

While it can be said that the work on the intermediate state has not produced any new and glaring inconsistencies, neither has it advanced our knowledge concerning the structure of this state appreciably. Shalnikov's experiments have produced indirect evidence of the existence of laminae, but there remains the discrepancy between his work and that of de Haas and Guinau regarding the diameter of the layers. Surface currents which delay the destruction of superconductivity, and the existence of which can hardly be doubted, remain completely unexplained. The existence of a surface energy, which influences the transition between the superconductive and the normal state, is strikingly demonstrated by the large delays in the magnetic induction of an aluminium sphere investigated by Shoenberg (1940a). Such effects, although smaller in magnitude, had been observed in very pure mercury and tin (Mendelssohn and Pontius, 1936b) and have been compared (Mendelssohn, 1935) with the supercooling of a very pure liquid below the melting point. By thermodynamic interpretation (London, H., 1935) of the electrodynamic equations of F. and H. London (1935), the surface tension responsible for the supercooling has been linked with the depth D to which a magnetic field can penetrate a superconductor and with the number of superconductive electrons.

§6. THE DEPTH OF PENETRATION

It is from experiments on quantities such as the depth D and the number of superconductive electrons that more information concerning the surface energy can be expected. Such an experiment of this kind is the determination of the eddy-current heating of a superconductor in an alternating field of high frequency carried out by H. London (1940). Assuming that, in addition to the superconductive electrons, there exist also electrons capable of carrying a normal current, eddy currents within D can be expected (London, H., 1934) if frequencies are employed at which the inertia of the electrons becomes a determining factor. Strictly speaking, the determination of the eddy-current heat is, therefore, an experiment on the normal and not on the superconductive electrons, but conclusions as to the number of superconductive electrons can be drawn if assumptions about the normal conductivity below T_c can be made. With a wave-length of 20·5 cm., heating effects were observed below T_c, indicating that the high-frequency resistance does not drop to the value zero at the transition point but shows a gradual decrease. This effect is attributed to a gradual increase in the number of superconductive electrons as the temperature is lowered. At absolute zero this number is estimated as 0·2 electrons per atom, while the experiments on thin mercury films (Appleyard, Bristow, London, H., and Misener, 1939) had yielded 0·05. Little can, however, be deduced, from these experiments, about the change with temperature of the number of superconductive electrons or of D. Shoenberg (1940b) has pointed out that his observations on superconductive colloids could not be brought in accord with London's penetration law, so that an independent determination of the variation of D with temperature, carried out by Casimir (1940), is of particular value. In his experiments the flux of induction is measured inside a coil wound closely on a container (~1·5 cm. in diameter) filled with mercury. While it is impossible

with this device to measure the absolute value of D, any variation of it will be recorded as a change in magnetic flux passing through the coil for a given change of the external magnetic field. The result (figure 6) was quite unexpected, and is in striking disagreement with the observations on thin films and colloids. It appears that there is *no* change in the value of D larger than 10^{-6} cm. between 3° and 3·9° K., and that a slight effect observed above 3·9° K. is due to spurious causes and not to a real change in D. In view of the disagreement with previous work, the author concludes that either the penetration law is incorrect or that the number of superconductive electrons depends on the dimensions of the superconductor. Casimir's experiments must necessarily render doubtful all previous conclusions regarding the value and significance of D not only on account of the result obtained but also because he was able to show that his results could be made to agree with the earlier work by introducing strain into the specimen. As a result of his work the whole conception of the penetration depth,

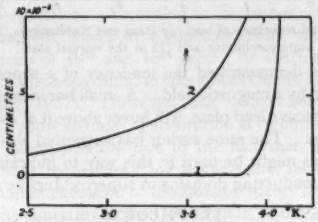


Figure 6. Variation of the depth of penetration with temperature : (1) Casimir (1940) and (2) Shoenberg (1940 b).

which forms an integral part of the London (1937) electrodynamics, has become unsatisfactory, and will certainly need revision.* An indication that the discrepancy in the observations may be a direct consequence of the uncertainty principle will be discussed below.

§7. MISCELLANEOUS EXPERIMENTS

De Haas and Rademakers (1940) have determined the thermal conductivity of lead in the normal and in the superconductive state. The results (figure 7) show that when superconductivity is suppressed by a field larger than H_c, the thermal resistance follows a smooth curve which is a continuation of the curve above the transition point. The superconductive phase, on the other hand, deviates sharply from this curve at T_c, showing a higher heat-resistance than the normal phase, which seems to indicate that the superconductive electrons do not take part in the heat-conduction process. The very strong rise of the thermal resistance in the neighbourhood of 2° K. would appear to lend support to the conclusion, arrived at from magnetic measurements (Daunt, Horseman and Mendelssohn, 1939), that near absolute zero there are, in the superconductor, no free electrons contributing to a Sommerfeld specific heat, and that all effectively

* An attempt to explain Casimir's result has recently been made by Ginsburg (1945).

free electrons in the metal have passed into the superconductive state at absolute zero.

Finally, two demonstration experiments should be mentioned. Justi (1942c) has used the high transition temperature of NbN to operate a small electro-magnet which is energized by a persistent current at the temperatures of liquid hydrogen.

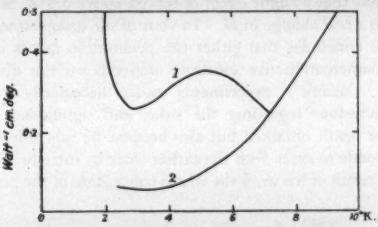


Figure 7. Thermal resistance of lead (de Haas and Rademakers, 1940): (1) in the superconductive and (2) in the normal state.

Arkadiev (1945a) has demonstrated the tendency of a superconductive surface to prevent penetration by a magnetic field. A small bar magnet of Fe-Ni-Al steel, when placed over a concave lead plate, will hover above it at a distance determined by its field and weight. The same author has suggested elsewhere (1945b) that small magnetic needles might be used in this way to indicate the existence and location of normally conducting domains in superconductive materials.

§8. THEORY

In a series of papers Koch (1940–43) has tried to develop a theory of superconductivity based on "classical" conceptions. He endeavours to substitute the notion of infinite conductivity for the fundamental equation of the London electrodynamics, which in his opinion is meaningless, but he finally arrives himself at the conclusion that his conceptions are insufficient to explain the magnetic behaviour of superconductors. Many of the author's statements (e.g. Meissner effect to be due to a temperature gradient; lack of data on the mechanism of the transition) can only be understood under the assumption that he is unfamiliar with much of the existing work on the subject. Some of his errors concerning the London theory, ordinary diamagnetism and infinite conductivity have been corrected by Neugebauer (1942). The magnetic properties of perfect conductors and superconductors have also been discussed by Mikura (1940). According to Stark (1943), the magnetic field of electrons should cause the formation of electron crystals (a conception which has been generally discarded as untenable), and he quotes this as proof for the ring structure of the electron advanced by him some years ago (Stark, 1937; Rao, 1938).

At the suggestion of v. Laue, Justi and Zickner (1941) have carried out a number of experiments on parallel superconductive coils which were coupled inductively. v. Laue (1941) concludes from the results of this work that there cannot exist a current-dependent energy which may be connected with the different degree of order between the superconductive and the normal metal.

If such an energy existed, the branch currents would not depend linearly on the total current, as was actually observed. In another paper the same author (v. Laue, 1942a) discusses the current distribution in circuits which are partly superconductive and partly normal.

The electrodynamical theory of superconductivity of F. and H. London has been extended by v. Laue (1942b, 43). In a relativistic interpretation the two fundamental equations of the London theory for the magnetic and the electric field are combined, while the explanations of the Meissner effect and the persistent currents remain unchanged. The parallel existence of two separate conduction processes, that of superconductivity, which obeys the London equations, and the process of normal conduction, obeying Ohm's law, are discussed. In a stationary field there exists only the supercurrent, and it is shown that the flux of induction through a multiply-connected superconductive body does not vary with time. If, on the other hand, the electric field varies with time, the process of normal conduction will, as had already been pointed out (London, H., 1934), come into operation. According to the author's calculations, the absorption of electromagnetic waves in a superconductor should become noticeable between frequencies of 10^{10} and 10^{14} sec^{-1}. At the upper limit of this range absorption has actually been found to take place (Daunt, Keeley and Mendelssohn, 1937; Hirschlaff, 1937). However, the above-mentioned experiments of H. London (1940) show an onset of absorption at frequencies as low as 10^{-9} sec^{-1}, the discrepancy being possibly due to the fact that the reduction of normal conductivity for high frequencies at low temperatures seems to have been neglected by v. Laue. Another extension of the London theory is the treatment of electric charges within the superconductor. A solution of the author's equations shows that undamped oscillations of the charge density, which are permitted in the London formulation of the theory, cannot in fact occur. The movement of a charge placed inside a superconductor, for instance by bombarding the metal with electrons, can be described, and it is shown that such an internal charge will move to the surface within a short time. In another paper the same author (v. Laue, 1942c) gives a general survey covering the position of the theoretical approach to the problem of superconductivity and discusses some of the existing discrepancies. He concludes that a superconductive sphere of macroscopic dimensions cannot be given an angular momentum by a stationary magnetic field, and that the gyro-magnetic experiment of Kikoin and Goobar (1940) utilizes the forces provided by an alternating field. However, it would seem that the frequency employed in these experiments was too low to justify this explanation.

Welker (1943) has discussed the processes which, according to electron-theoretical conceptions, are likely to furnish an energy of the order of kT_e needed to reverse the velocity vector of an electron in an assembly of electrons which is treated like a liquid. The velocity-independent Coulomb force is not suitable to provide this energy, by which, in any case, T_e would come out 10^4 times too high. The magnetic interaction, which is of the order of

$$v^2e^2/c^2d,$$

where v is the zero-point velocity and d the average distance between electrons,

furnishes an energy of the correct order ($\sim 1{\cdot}8 \times 10^{-16}$ erg), which leads to a transition point at about 1°K. However, the classical magnetic interaction, which gives a repulsion of antiparallel currents, does not lead to the required distance function in which neighbouring electrons ought to have opposite velocities. An attraction between antiparallel currents, on the other hand, results from the magnetic exchange interaction between electrons of equal spin, and it is this force which, according to the author, must be responsible for the small energy kT_c. The theory rests on the assumption, made plausible in an analogy to crystal structure, that in certain circumstances this exchange interaction can outweigh the classical magnetic repulsion force. In a paper on this theory Sommerfeld (1941) states that in calculations, which have not been published, Christlein has derived the form of the threshold curve on the basis of Welker's assumptions. This calculated curve has, at low temperatures, the form of a parabola from which it deviates however near T_c, so that its tangent is normal to the T-axis at T_c. The latter feature is certainly not in agreement with the observed H_c/T-curves which, in first approximation, have an undisturbed parabolic form. Sommerfeld points out that Welker's formula, relating the threshold field at absolute zero H_0 and the electronic specific heat γ with T_c,

$$\frac{H_0\sqrt{V}}{T_c\sqrt{\gamma}} = \text{constant}$$

(V being the atomic volume), can also be derived from the thermodynamical considerations, determining the constant as $\sqrt{2\pi}$, a derivation which had already been given earlier (Daunt, Horseman and Mendelssohn, 1939).* In order to account for deviations from the parabolic form, Sommerfeld introduces the generalized version,

$$\frac{H_0\sqrt{V}}{T_c\sqrt{\gamma}} = \sqrt{\frac{2\pi}{\alpha}},$$

but considers appreciable deviations of α from 1 as contrary to the theory. Welker (1943), on the other hand, postulates the value 0·58 for α, which is only approached in indium ($\alpha = 0{\cdot}63$), while mercury ($\alpha = 0{\cdot}82$) and thallium ($\alpha = 1{\cdot}04$) do not fit the equation. The same will be true for tin and lead, whose H_c/T-curves are approximately parabolic ($\alpha = 1$).

Ariyama (1942) has introduced the magnetic interaction of the orbital momenta of the electrons into the Bloch-Hartree model, and from this he deduces magnetic interaction in superconductors at a distance of 10^{-6} cm.

Daunt and Mendelssohn (1942) have drawn attention to a striking analogy between superconductivity and the λ-phenomenon of liquid helium, in both cases frictionless transport being considered the fundamental phenomenon. From the observation of zero Thomson-heat it is concluded (Daunt and Mendelssohn, 1946 b) that the superconductive electrons have zero entropy,

* In this paper the constant was given as $\sqrt{2\pi/V}$ because it had been overlooked that γ in Welker's original communication referred to the specific heat per cc. The author is indebted to Dr. J. A. Kok for drawing his attention to this error.

and that the thermal states of the electronic energy spectrum are separated from the set of ground states by a gap. The superconductive electrons, remaining energetically at absolute zero even at finite temperatures, have *no* specific heat, the observed anomaly being due to a lifting of electrons over the gap. The super-currents are considered to derive their momentum from zero-point energy (Mendelssohn, 1945), which would be in accordance with the observation that these currents are not accelerated by an electric field. The variation of the critical rate of frictionless transport, as expressed by H_c, is taken to indicate the number of superconductive electrons. The hypothesis has been taken up by F. London (1945), who has made an estimate of the temperature variation of the number of superconductive electrons n, and has pointed out that the critical rate of transport is of the order of $nh/2\pi$. Daunt and Mendelssohn (1946 a) have shown that when the rate of the frictionless flow is integrated over the cross-section of a superconductor of diameter d the relation can be expressed as

$$v \sim h/md,$$

where v is the average velocity of the superconductive electrons and m their mass. This is, of course, the zero-point velocity conferred by indeterminacy on a particle which is confined to a space of the diameter of the order of d. Thereby the conception of zero-point energy* as the basis of supercurrents appears to be considerably strengthened. The relation between the average velocity and the linear dimension of the superconductor is particularly interesting. It is fully born out by the well-known fact that the supercurrent is proportional to the *diameter* of the wire carrying it, which has always been explained as being due to surface currents. It now appears that this phenomenon may be a direct consequence of indeterminacy which does not permit the localization of a superconductive electron within narrower limits than the dimensions of the superconductor. Superconductivity thus appears as a phenomenon in which the quantum constant becomes the determining factor in systems of macroscopic size. As a result of this impossibility to localize electrons, currents in a superconductor must appear as surface currents, but whether any physical significance has to be ascribed to the depth of penetration, as far as the superconductive electrons are concerned, is not clear. It is therefore possible that the discrepancy between the observations of Casimir (1940) and the earlier experiments (Appleyard, Bristow, London, H., and Misener, 1939; Shoenberg, D., 1940 b), is due to the fact that this quantity has been given too literal an interpretation.

REFERENCES

Alekseyevsky, N., 1940. *J. Phys. U.S.S.R.*, **3**, 443.
Alekseyevsky, N., 1945 a. *J. Phys. U.S.S.R.*, **9**, 147.
Alekseyevsky, N., 1945 b. *J. Phys. U.S.S.R.*, **9**, 217.
Alekseyevsky, N., 1945 c. *J. Phys. U.S.S.R.*, **9**, 350.
Appleyard, E. T. S., Bristow, J. R., London, H. and Misener, A. D., 1939. *Proc. Roy. Soc.*, A., **172**, 540.

* This " zero-point energy of mean free path " (London, F., 1939) which arises out of indeterminacy must not be confused with the zero-point energy of the free electrons due to Fermi-Dirac statistics. It exists in a Bose-Einstein gas as well as in a Fermi-Dirac gas, provided that the particles have finite cross-section.

ARIYAMA, K., 1942. *Z. Phys.*, **119**, 174.
ARKADIEV, V. K., 1945 a. *J. Phys. U.S.S.R.*, **9**, 148.
ARKADIEV, V. K., 1945 b. *C.R. Acad. Sci. U.S.S.R.*, **47**, 15.
ASCHERMANN, G., FRIEDRICH, E., JUSTI, E. and KRAMER, J., 1941. *Phys. Z.*, **42**, 349.
ASCHERMANN, G. and JUSTI, E., 1942. *Phys. Z.*, **43**, 207.
BECKER, R., HELLER, G. and SAUTER, F., 1933. *Z. Phys.*, **85**, 772.
BRUCKSCH, W. F. and ZIEGLER, W. T., 1942. *Phys. Rev.*, **62**, 348.
CASIMIR, H., 1940. *Physica*, **7**, 887.
CASIMIR, H., 1943. *Physica*, **10**, 757.
DAUNT, J. G., HORSEMAN, A. and MENDELSSOHN, K., 1939. *Phil. Mag.*, **27**, 754.
DAUNT, J. G., KEELEY, T. C. and MENDELSSOHN, K., 1937. *Phil. Mag.*, **23**, 264.
DAUNT, J. G. and MENDELSSOHN, K., 1937. *Proc. Roy. Soc.*, A, **160**, 127.
DAUNT, J. G. and MENDELSSOHN, K., 1938. *Nature, Lond.*, **141**, 116.
DAUNT, J. G. and MENDELSSOHN, K., 1942. *Nature, Lond.*, **150**, 604.
DAUNT, J. G. and MENDELSSOHN, K., 1946 a. *Phys. Rev.*, **69**, 126.
DAUNT, J. G. and MENDELSSOHN, K., 1946 b. *Proc. Roy. Soc.*, A, **185**, 225.
GINSBURG, V., 1945. *J. Phys. U.S.S.R.*, **9**, 305.
GORTER, C. J. and CASIMIR, H., 1933. *Physica*, **1**, 306.
DE HAAS, W. J. and GUINAU, O. A., 1936. *Physica*, **3**, 534.
DE HAAS, W. J. and RADEMAKERS, A., 1940. *Physica*, **7**, 992.
DE HAAS, W. J. and VOOGD, J., 1931. *Proc. Acad. Sci., Amst.*, **34**, 192 (*Commun. Phys. Lab. Univ. Leiden*, 214 c).
HIRSCHLAFF, E., 1937. *Proc. Camb. Phil. Soc.*, **33**, 140.
JACKSON, L. C., 1939. *Rep. Prog. Phys.*, **6**, 335.
JUSTI, E., 1941. *Phys. Z.*, **42**, 325.
JUSTI, E., 1942 a. *Phys. Z.*, **43**, 130.
JUSTI, E., 1942 b. *Ann. Phys., Lpz.*, **42**, 84.
JUSTI, E., 1942 c. *Elektrotech. Z.*, **63**, 577.
JUSTI, E., 1943. *Phys. Z.*, **44**, 469.
JUSTI, E. and ZICKNER, G., 1941. *Phys. Z.*, **42**, 257.
KAMERLINGH ONNES, H. and TUYN, W., 1922. *Versl. Gewone Vergad. Akad. Amst.*, **31**, 475 (*Commun. Phys. Lab. Univ. Leiden*, 160 b).
KEESOM, W. H. and DÉSIRANT, M., 1941. *Physica*, **8**, 273.
KEESOM, W. H. and VAN LAER, P. H., 1938. *Physica*, **5**, 193.
KIKOIN, I. K. and GOOBAR, S. W., 1940. *J. Phys. U.S.S.R.*, **3**, 333.
KOCH, K. M., 1940. *Z. Phys.*, **116**, 586.
KOCH, K. M., 1941. *Z. Phys.*, **118**, 1.
KOCH, K. M., 1942. *Z. Phys.*, **120**, 86.
KOCH, K. M., 1943. *Z. Phys.*, **121**, 488.
VAN LAER, P. H. and GROENEWOLD, K., 1939. *Physica*, **6**, 473.
LANDAU, L., 1937. *Sow. Phys.*, **11**, 129.
LANDAU, L., 1938. *Nature, Lond.*, **147**, 688.
LANDAU, L., 1943. *J. Phys. U.S.S.R.*, **7**, 99.
LASAREW, B. and GALKIN, A., 1944. *J. Phys. U.S.S.R.*, **8**, 376.
LASAREW, B. and KAN, L., 1944 a. *J. Phys. U.S.S.R.*, **8**, 193.
LASAREW, B. and KAN, L., 1944 b. *J. Phys. U.S.S.R.*, **8**, 361.
LASAREW, B. and NAKHUTIN, J. E., 1942. *J. Phys. U.S.S.R.*, **6**, 116.
v. LAUE, M., 1941. *Z. Phys.*, **118**, 455.
v. LAUE, M., 1942 a. *Z. Phys.*, **120**, 578.
v. LAUE, M., 1942 b. *Ann. Phys., Lpz.*, **42**, 65.
v. LAUE, M., 1942 c. *Phys. Z.*, **43**, 274.
v. LAUE, M., 1943. *Ann. Phys., Lpz.*, **43**, 233.
LONDON, F., 1937. *Une conception nouvelle de la supraconductibilité* (Paris : Hermann et Cie.).
LONDON, F., 1939. *J. Phys. Chem.*, **43**, 49.
LONDON, F., 1945. *Rev. Mod. Phys.*, **17**, 310.
LONDON, F. and H., 1935. *Proc. Roy. Soc.*, A, **149**, 71.
LONDON, H., 1934. *Nature, Lond.*, **133**, 447.

LONDON, H., 1935. *Proc. Roy. Soc.*, A, **152**, 650.
LONDON, H., 1940. *Proc. Roy. Soc.*, A, **176**, 522.
MEISSNER, W., 1928. *Phys. Z.*, **29**, 897.
MEISSNER, W., 1930. *Z. Phys.*, **60**, 181.
MEISSNER, W., FRANZ, H. and WESTERHOFF, H., 1932. *Z. Phys.*, **75**, 521.
MEISSNER, W. and OCHSENFELD, R., 1933. *Naturwissenschaften*, **21**, 787.
MENDELSSOHN, K., 1935. *Proc. Roy. Soc.*, A, **152**, 34.
MENDELSSOHN, K., 1941. *Nature, Lond.*, **148**, 316.
MENDELSSOHN, K., 1945. *Proc. Phys. Soc.*, **57**, 371.
MENDELSSOHN, K. and BABBITT, J. D., 1935. *Proc. Roy. Soc.*, A, **151**, 316.
MENDELSSOHN, K. and MOORE, J. R., 1935. *Proc. Roy. Soc.*, A, **151**, 334.
MENDELSSOHN, K. and MOORE, J. R., 1936. *Phil. Mag.*, **21**, 532.
MENDELSSOHN, K. and PONTIUS, R. B., 1936 a. *Physica*, **3**, 327.
MENDELSSOHN, K. and PONTIUS, R. B., 1936 b. *Nature, Lond.*, **138**, 29.
MIKURA, Z., 1940. *Proc. Phys.-Math. Soc. Japan*, **22**, 733.
NAKHUTIN, J. E., 1939. *J. Exp. Theor. Phys. U.S.S.R.*, **9**, 708.
NAKHUTIN, J. E., 1942. *J. Phys. U.S.S.R.*, **6**, 114.
NEUGEBAUER, T., 1942. *Z. Phys.*, **119**, 581.
RAO, B. S. MADHAVA, 1938. *Phys. Z.*, **39**, 187.
SHALNIKOV, A., 1942. *J. Phys. U.S.S.R.*, **6**, 53.
SHALNIKOV, A., 1945. *J. Phys. U.S.S.R.*, **9**, 202.
SHOENBERG, D., 1935. *Proc. Roy. Soc.*, A, **152**, 10.
SHOENBERG, D., 1940 a. *Proc. Camb. Phil. Soc.*, **36**, 84.
SHOENBERG, D., 1940 b. *Proc. Roy. Soc.*, A, **175**, 49.
SHUBNIKOV, L., CHOTKEVITCH, W. J., SHEPELEV, J. D. and RJABININ, J. N., 1936. *Sow. Phys., Sondernummer*, June, pp. 46/47.
SILSBEE, F. B., 1916. *J. Wash. Acad. Sci.*, **6**, 597.
SIZOO, G. J. and KAMERLINGH ONNES, H., 1925. *Versl. Gewone Vergad. Akad. Amst.*, **34**, 606 (*Commun. Phys. Lab. Univ. Leiden*, 180 b).
SOMMERFELD, A., 1941. *Z. Phys.*, **118**, 467.
STARK, J., 1937. *Phys. Z.*, **38**, 269.
STARK, J., 1943. *Phys. Z.*, **44**, 211.
STEINER, K. and GERSCHLAUER, W., 1942. *Ann. Phys., Lpz.*, **42**, 98.
STEINER, K. and SCHOENECK, H., 1943 a. *Phys. Z.*, **44**, 341.
STEINER, K. and SCHOENECK, H., 1943 b. *Phys. Z.*, **44**, 346.
TUYN, W. and KAMERLINGH ONNES, H., 1926. *J. Franklin Inst.*, **201**, 379 (*Commun. Phys. Lab. Univ. Leiden*, 174 a).
WELKER, H., 1943. *Phys. Z.*, **44**, 134.
ZIEGLER, W. T., BRUCKSCH, W. F. and HICKMAN, J. W., 1942. *Phys. Rev.*, **62**, 354.

SPECTROSCOPY APPLIED TO MOLECULAR PHYSICS

By J. CABANNES,
Faculté des Sciences de Paris

§1. INTRODUCTION

If the atoms in a molecule were to move at random, as do the molecules themselves in a drop of water, to imagine the existence of isomers with different properties would be inconceivable. A molecule is characterized not only by the nature and number of its constituent atoms but also by the positions they occupy, and it is the purpose of stereo-chemistry to determine the arrangement of atoms that will explain their chemical reactions. On their own resources the chemists have gone a long way in the determination of molecular structures, and their methods of analysis and synthesis still play the major rôle in the case of large molecules, but there is always an element of doubt as to the certainty of the conclusions which can be drawn from the reactions. Physical methods are more accurate and far-reaching and avoid the destruction of the molecules. Instead of the more drastic chemical actions, the gentler physical perturbations are employed, and the molecules are observed, so to speak, without touching them.

On the other hand internal re-arrangements, such as the change of the ketonic form of ethyl acetoacetate ether to the enolic at room temperature, or the transformation of ammonium isocyanate into urea on heating, reveal intramolecular movements which become more pronounced as the temperature is raised A single molecule may have several stable configurations, and transition from one to the other takes place with varying degrees of facility. If the potential barrier is low, chemistry can neither isolate nor detect these different structures; such is the case with molecules in which free rotation is possible about certain bonds. The physicist, however, can identify the molecular isomers in a mixture and also determine their respective configurations without having to separate them.

Two physical methods have proved particularly useful. One employs the action of molecules upon radiation of very short wave-length (e.g. x rays or electrons) and the other their action on light.

The diffraction of an x-ray or electron beam by gas molecules is a theoretically perfect method of analysing molecular structures. The distance between the scattering centres in the molecule follows from the distance between the interference maxima, but very few results have been obtained by this method. Its drawback is that it is applicable only to gases and vapours; in fact, x-ray and electron diffraction in liquids serves to show the regularities produced in a dense medium by the arrangement of the molecules rather than the atomic configuration in the molecules.

Accordingly the most numerous and most recent successes of molecular physics have been obtained in research on the scattering of light and on the selective absorption of infra-red radiation by which, in many problems, the liquid phase may be studied instead of the much more difficult gas phase. As it is obvious that x-ray diffraction, in which the wave-lengths are of the same order as the interatomic distances, may reflect the details of molecular structure, at first sight it appears paradoxical that wave-lengths a thousand times greater can be used effectively. But it is in this region where the intra-molecular movements, having frequencies which fall within the optical region, can interact with radiation. In particular the rotation of the molecules and the vibrations of the nuclei in each of them determine the infra-red absorption and the Raman spectra. If the frequency and direction of these movements are known from spectroscopic analysis, the arrangement of the nuclei in the molecule and the field of force that maintains them in it can be deduced.

To sum up, there is a two-fold problem to be solved: (i) To determine the arrangement of the nuclei and, more particularly, the symmetry of this arrangement; (ii) to calculate the field of force which maintains them in their equilibrium position. To obtain the maximum information, it is necessary to know both the infra-red absorption spectrum (if possible with the structure of the bands) and the scattered spectrum (with the intensity and depolarization of the Raman lines).

Even if we limit ourselves to the most recent works, it will be impossible to make a critical study of all the structures that have been published. Therefore this *Report* will be limited to a review of the progress made in the application of spectroscopic technique to molecular physics and to a few typical examples to illustrate the methods and the value of the results obtained in the determination of molecular structures by spectral analysis.

PART I.—TECHNIQUES

§ 2. PROGRESS IN THE STUDY OF RAMAN SPECTRA

(i) *Characteristics of modern spectrographs*

The efficiency of a spectrograph depends essentially upon its resolving power and the intensity of the image. By 1927 gratings and interferometers had already reached a high degree of perfection; they rendered possible the separation of very close spectral lines and the determination of their width. But the spectrographs then in use, including the prism instruments, gave weak intensities and were not adapted to the analysis of the feeble lines of molecular scattering. It was necessary to multiply the illumination, E, on the photographic plate by 10, or even 100, to obtain a sufficiently dense negative in a practicable time of exposure, t. The empirical law of Schwartzschild ($Et^{\gamma} = \text{constant}$) shows the gain to be obtained by raising the illumination on the sensitive plate as much as possible. The value of γ is usually 0·7 to 0·9; when it is 0·8 the time of exposure is decreased by a factor of 300 if the illumination is increased by a factor of 100.

From geometrical considerations the intensity of illumination of a point on a photographic plate depends only on the aperture of the objective. It is practically proportional to $\sin^2 \frac{u}{2}$, where u is the angle of the cone of rays leaving the objective. The solution is then immediately apparent: employ lenses of the largest possible diameter, reduce the focal length of the objective and, in order not to reduce the dispersion, use two or even three prisms. The principal difficulty is the design of a lens of large numerical aperture and large dimensions which will give a good spectral image on a flat photographic plate.

There are already commercial achromatic lenses with apertures in the neighbourhood of F/1 available, but the necessity for having a sufficiently linear dispersion compels one to provide for greater dimensions. Reproduction of existing photographic lenses on a larger scale is unsatisfactory because aberration increases simultaneously and nothing is gained. What is required is an increase in the size of the image without a decrease in its definition. This is hard to achieve with an achromatic objective, as the dispersion curves of glasses needed for achromatism differ too widely in order to correct spherical aberration over a wide spectral range. But one is not concerned with achromatism when designing a lens for a spectrograph, therefore it is possible to choose glasses with the same order of dispersion, and the objective, corrected for a mean wave-length, remains corrected throughout the spectral range. Naturally it is necessary to incline the sensitive plate to the axis, but this is a minor inconvenience.

On these principles Cojan has constructed spectrographs with which the Raman effect has been studied in France since 1928.

Table 1 gives the characteristics of the most recent models.

Table 1

Prisms	Aperture	Dispersion
I. 3 prisms : dense flint	D=F/8 =140 mm.	5–10 A./mm. in the interval 4100–4700 A.
II. 1 prism : very dense flint	D=F/0·7=110 mm.	150–900 A./mm. in the interval 4340–6500 A.
III. 2 prisms : dense flint	D=F/1·9=320 mm.	20–35 A./mm. in the interval 4100–4700 A.
IV. 3 prisms : flint	D=F/5·6=107 mm.	8–15 A./mm. in the interval 4100–4700 A.
V. 3 prisms : dense flint	D=F/2·3= 90 mm.	20–40 A./mm. in the interval 4100–4700 A.

The prisms of dense flint have an index μ_D and a constringence ν of about 1·62 and 36·5 respectively; those of the very dense flint are 1·665 and 33.

Model I, having weak intensity and high dispersion, was used for the study of rotation lines in gases and of lines produced by the oscillation of molecules in organic crystals (Kastler and Rousset, 1941); but exposures lasted 12 days (in oxygen at atmospheric pressure).

Model II, on the other hand, gives high intensity but low dispersion. However, owing to the quality of the image given by Cojan's objective, the lines of iron 4459–4462 A. and 5447–5455 A. can be separated on standard negatives. The longitudinal aberration, at a minimum and equal to F/4000 in the region of 4861 A., does not exceed F/800 at the limits of the range 4340-6500 A. We have made this spectrograph for analysing the light of the night sky in which we have found bands of nitrogen belonging to the Vegard-Kaplan system (Cabannes and Dufay, 1935); we have subsequently used it to measure the depolarization of Raman lines by ordinary gases under atmospheric pressure (Cabannes and Rousset, 1940). By replacing Cojan's objective by an achromatic objective made by Rayton, which is less well corrected but of larger aperture (F/0·59), we were able to obtain the principal lines of liquid benzene in one second and three Raman lines of hydrogen at atmospheric pressure in 10 minutes. After one hour the Raman lines of the oxygen and nitrogen in the air of the laboratory become apparent.*

In the third model Cojan has simultaneously increased the intensity of the first and the dispersion of the second by using an objective of very large diameter. As we have shown above, this was a difficult problem. Nevertheless, the image of the spectrum, flat to within $\pm 0{\cdot}2$ mm. in the region 4100–4700 A., is as good as that in ordinary apparatus; two lines of iron 1·5 A. apart at 4920 A. were separated. This spectrograph gives the Raman spectrum of liquid carbon disulphide in 2 minutes. It is at present exhibited in the Palais de la Découverte in Paris.

Models IV and V, of smaller dimensions, are in use, one in the chemical laboratory of the Faculté des Sciences de Lille, the other in the physical research laboratory in the Sorbonne.

(ii) *Photoelectric detection*

Photographic plates are employed in all the preceding apparatus, and the exposure, development and microphotometric recording of the negative takes up much time. To eliminate this inconvenience, Rank, Pfister and Coleman (1942), Rank, Pfister and Grimm (1943), and Dieke and Crosswhite (1945) used photo-electric detection, rendered possible by the progress made in the construction of electron-multipliers in the United States. Raman lines, which have been separated by a grating-type of monochromator, fall on an electron-multiplier (R.C.A. 931) and the response is recorded. The initial current is amplified by 10^8. The sensitivity is limited only by the thermionic emission at the surrounding temperature. With this new apparatus lines can be observed of just as low an intensity as that detectable by a photographic plate.

(iii) *Methods for the analysis of light scattered by powders*

The study of the Raman effect in a liquid, gas or single crystal is no longer difficult, but the need to study solid bodies and the impossibility of obtaining the majority of them in sufficiently large crystals has forced physicists to devise methods for the spectoscopy of crystalline powders.

* With a Hilger spectrograph of high intensity (High Speed Hilger E.518 Glass Spectrograph) Cleveland (1943) obtained eight Raman lines of CCl_4 in 1 minute.

The powdered substance can be observed either by transmission or by reflection; but in both cases the incident light is reflected or refracted by the grains with such an intensity that most of the Raman lines are masked on the photographic plate by an intense continuous background, unless two essential conditions are strictly adhered to, namely : (i) monochromatic light must be used; (ii) the exciting radiation in the re-emitted light must be eliminated.

Ananthakrishnan was the first person to obtain, in 1936, spectra of crystalline powders. He put a coloured filter in the incident beam and isolated, as well as possible, one of the visible lines of mercury, which was subsequently absorbed in a complementary filter placed in the re-emitted beam. The efficiency of the method depends on the choice of filters, and some progress has been made in this direction in recent years.

Conrad-Billroth, Kohlrausch and Reitz (1937) and Reitz (1940) have obtained better results by substituting two dispersion systems for the coloured filters. A double monochromator selects one line of the mercury arc ($\lambda = 4358$ A.) with which the surface AB of the crystalline powder is illuminated; a second dispersion unit receives the light re-emitted by the powder and projects a series of monochromatic images of the surface AB in the plane P (figure 1). Among

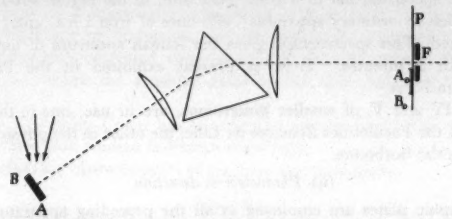


Figure 1.

these images is one, A_0B_0, given by the line 4358 A. The slit F of the spectrograph (to detect the Raman spectrum) is placed in the screen P beyond the image A_0B_0. In this way unwanted light is avoided in so far as the scattering of light by the prism and lenses of the second dispersion unit can be neglected.

Each half of the double monochromator consists of a prism of ethyl cinnamate, and the temperature is kept constant to within half a degree. A condensing lens covers the intermediate slit. A very reduced image of the second prism face, from which the light emerges, is projected on to the powder to be studied. The essential part of the second dispersion unit is a 60° flint prism. The principal spectrograph, made by Zeiss, has three prisms; use is made of one of two objective lenses (apertures F/4·5 and F/2, focal lengths 27 cm. and 11 cm.), according to whether greater dispersion or greater intensity is required.

The spread of the continuous background, which extends towards the lower frequencies from the 4358 line, varies according to the width of the intermediate slit of the double monochromator; it can be reduced to 170 cm.$^{-1}$. In general it is limited to 400 cm.$^{-1}$, and the exposures have a duration of 30 to 40 hours.

In the method given above the powder was examined by reflection; Menzies and Mills (1935) carried out and described observations by transmission. A very simple technique which has not been described yet has been used recently in the Physical Research Laboratory at the Sorbonne. A spectrograph of type II (table 1) is used as a monochromator. In the plane of the spectrum a diaphragm isolates line 4358 of the mercury arc, and the powder to be examined is placed against it. The image of the powder is projected on to the slit of a spectrograph (type V) without interposing a coloured filter. The only difficulty is in the choice of the thickness of the sample according to its transparency and crystal size. Thus, in a few hours, all the Raman lines of frequency equal to, or greater than, 250 $cm.^{-1}$ can be obtained for a substance such as salol. This method allows of the study of coloured powders which selectively absorb the exciting radiation.

§ 3. PROGRESS IN THE STUDY OF INFRA-RED ABSORPTION SPECTRA

(i) *Detection*

Absorption bands produced by the fundamental vibrations of nuclei have wave-lengths greater than $2{\cdot}7\,\mu$; in this region heat detectors are almost the only instruments applicable, and the use of the thermopile has become universal. Progress has followed the improvements to galvanometers, and it seems that the limit of sensitivity has been reached. The moving frame of the galvanometer is subject to unavoidable Brownian agitation due to bombardment by air molecules and to thermal agitation of the electrons in the circuit. This movement of the frame limits the possibility of detecting constant currents. One can hardly hope for better than the galvanometers which have already been made, and which are sensitive to Brownian movement. The best use must be made of the existing instruments.

The problem is to amplify the changes produced by the continuous current from the thermopile due to the incident radiation, without also amplifying the Brownian movement or the drifts which result from the fluctuations in temperature in the thermopile produced by exchanges of heat with its surroundings. In principle, one can eliminate the drifts by periodically interrupting the incident radiation to be analysed, whilst the spectrum passes over the detector, so that the displacement of the zero can be followed and allowance made for it. With a thermo-electric junction of low heat-capacity it is possible to attain 10 interruptions per second and to amplify the current modulated in this way.

In Pfund's radiometer, the theory of which was formulated by Hardy (1930), the interruptions have a frequency equal to the natural frequency of the somewhat damped galvanometer G_1 (amplifying galvanometer). Under these conditions the amplitude of the alternating deflection of G_1 is 20 times greater than the fixed deflection which would be given by uninterrupted radiations; this reduces perturbations (drifts and Brownian movements) which are not amplified. The spot of the galvanometer G_1 in its turn illuminates a thermopile through an aperture cut out in such a way that the energy falling on the thermopile is proportional to the deflection of G_1. In the circuit of the second pile is placed

a galvanometer G_2 (indicating galvanometer), the deflections of which are 20 times as great as those of G_1. Thus a total amplification of 400 is obtained and disturbances are practically eliminated. However, 30 seconds must be allowed for the radiometer to reach equilibrium, and, with this limitation, it is useless for spectroscopy.

Firestone (1932) made a periodic radiometer, on the same principles, in which the time constant was reduced to 6 seconds with very little diminution in the sensitivity of the detector; this time constant is long enough to eliminate Brownian movement without increasing too greatly the time required to record spectra. At the same time, to ensure a rapid response the two galvanometers G_1 and G_2 had the same frequency, and were adjusted for critical damping. Lastly, the galvanometer spot is used to illuminate not a thermopile but a photo-electric cell, so that the current can be amplified by means of a valve before being passed to the indicating galvanometer.

In spite of the perfection attained by these thermoelectric detectors, it is necessary to keep in mind the fact that the thermometric method is questionable, and that the maximum efficiency is extremely small, The thermoelectric receivers operate between room temperature T, which can be regarded as constant, and the temperature $T+\Delta T$ of the absorbing surface, so that the maximum yield is $\Delta T/T$. If ΔT has a value of about 10^{-7}, the efficiency is 3×10^{-10}. Whatever type of receiver is imagined, the electric forces created will still be very feeble and the appropriate galvanometers for measuring them will have to be delicate. Consequently, in spite of the very great sensitivity of electrical methods, it may be better to observe directly the flow of available heat. This involves determination of the increase in the mean kinetic energy of the molecules. Evaporation furnishes a simple and efficacious method, which has already given interesting results (Czerny and Mollet, 1937–38).

In all cases the receiving surface must be perfectly absorbent for the radiations being studied, and as reflective as possible for those outside this region. In the near infra-red ($\lambda<20\mu$) there are several satisfactory solutions to this problem: since carbon and platinum black become less absorbent as wave-length increases, they are mixed with sodium silicate. Pfund obtained very absorbent receivers by condensing a thin layer of metal, which had been evaporated under reduced pressure, on the thermopile junction. But to examine the far infra-red ($\lambda>20\mu$) the intense radiations of short wave-length emitted by different parts of the spectrograph must be eliminated. If the surface is coated with thallium chloride distilled in vacuum, only radiations in the region $30–130\mu$ are absorbed. The surface acquires the double property of being insensitive to unwanted radiations and of emitting none of the radiations in the region $5–10\mu$, which allows it to reach a higher temperature.

(ii) *Dispersion in the near and far infra-red*

Spectrographs made with prisms, or prisms and gratings combined, have been in use since 1920, and have often been described (Coblentz, 1920; Lecomte, 1929–36). The most recent progress is in the production of large artificial crystals from which can be cut plates and prisms transparent to about

30 μ (LiF, NaF, NaCl, KCl, KBr, KI). According to the spectral regions being observed, the material of the prism is chosen to give the maximum separation. In practice lines 1 cm.$^{-1}$ apart in the neighbourhood of 3 μ and 0·5 cm.$^{-1}$ in the region 10 μ are separated; the positions are accurate to 0·05 cm.$^{-1}$.

However, analysis of molecular spectra necessitates the exploration of the far infra-red, where difficulty is encountered in the lack of transparency of the prisms and the weak intensity of the radiation. If the energy of black-body radiation of 1·22 μ at 2360° A. is taken as unity, that at 70 μ is less than two millionths. If the shape of the grating elements is such that the observed radiations are concentrated into one spectrum, a reflection grating can be used up to 400 μ. Spectrographs for far infra-red have been improving for 20 years (1920–1940).

In 1934 Barnes, with two mirrors of 15 cm. diameter and a wire grating which eliminated spectra of even order, obtained, between 70 and 180 μ, deflections of several millimetres, but the breadth of the slit limited the resolution. At 50 μ Barnes separated lines 4·42 cm.$^{-1}$ apart, at 100 μ lines 1·67 cm.$^{-1}$ apart, and at 150 μ lines 1·14 cm.$^{-1}$ apart. The position of the absorption maxima was found with an accuracy of 1 cm.$^{-1}$. The doubling of the rotation lines of ammonia by the tunnel effect, predicted by Dennison and Hardy (1932), did not show clearly in his spectra.

In order to increase the resolution, Randall (1938) used a mirror of much larger dimensions (diameter 60 cm.; focal length 90 cm.) in which spherical aberration had been reduced to a minimum. The sensitivity was increased with Firestone's amplifier. He obtained greater intensity of the spectrum by using Wood and Trowbridge's "echelette" grating constructed under Barker's direction, instead of wire or laminated gratings. In this way a 2-mm. slit gives deflections from 350 mm. (between 64 and 62 μ) to 100 mm. (near 73 μ) with a resolving power of 0·5 cm.$^{-1}$. The absorption maxima are determined with an accuracy of about 0·01 cm.$^{-1}$. Interesting records for the rotation lines of ammonia obtained in this way show that they are doublets. The distance between the two components for $J = 5, 6, 7$ was $1{\cdot}35 \pm 0{\cdot}03$ cm.$^{-1}$ (Wright and Randall, 1933) and fine structure was observed for values of $J = 11, 12, 13, 14, 15$ (Foley and Randall, 1941). The far more complicated rotation spectrum of water has been completely analysed from 18 to 135 μ by Randall, Dennison, Ginsburg and Weber (1937). More recently Hopf (1940) has measured new absorption maxima between 140 and 400 μ, using a large slit which isolated a spectral region of 18 μ.

To recapitulate, throughout the whole infra-red region up to 140 μ the same resolution of 0·5 cm.$^{-1}$ is obtained. Most of the rotation frequencies of molecules are found in this region, and can probably be measured with greater precision in infra-red spectra than in Raman spectra. Pure rotation lines, which are generally difficult to separate in Raman spectra, are now recognizable in the far infra-red. However, if the molecule contains heavy atoms the most intense rotation lines are found beyond 150 μ in a still unexplored region.

Generally, if the frequencies become too low (molecular rotation, vibrations of heavy nuclei, intermolecular reactions), their determination by all methods is difficult; in the far infra-red because the intensity of the radiation is inadequate,

and in Raman spectra because the corresponding lines are obscured by the light emitted without a change of wave-length.

(iii) *Analysis of light absorbed by powders*

The analysis of light absorbed by gases alone, in which the bands are resolvable, gives complete information about molecular structure. But many slightly volatile substances, or substances which are decomposed on heating, cannot be observed in the vapour state; others, which are insoluble in transparent liquids, have a high melting point and do not give large enough crystals to be cut into plates, can only be observed as powders. In the latter case Lecomte and his colleagues have systematically employed a method which is applicable to very many organic crystals. By placing the finely powdered substance between two transparent plates, they obtained good absorption spectra in the whole infra-red range except in the region near the visible. Since the wave-length of the radiation used was large, its scattering by the powdered medium was not strong enough to mask the selective absorption bands. Good absorption spectra are obtained if the thickness of the liquid or powder is between a few microns and a few tenths of a millimetre.

PART 2. VIBRATION OF NUCLEI IN POLYATOMIC MOLECULES

§4. THEORY OF SMALL VIBRATIONS OF NUCLEI IN POLYATOMIC MOLECULES

(i) *Definition of fundamental frequencies*

Every molecule is a more or less complex arrangement of atomic nuclei and electrons. The state of the system at each instant is given by the function ψ of the coordinates $X_1 \ldots Z_p$ of the nuclei and the coordinates $x_1 \ldots z_n$ of the electrons, from which the probability of finding the system in the small volume $d\tau = dX_1 \ldots dZ_p dx_1 \ldots dz_n$ of the spatial configuration at any given instant is $\psi\psi^* d\tau$.

As the nuclei have masses thousands of times greater than that of the electrons, their deflection due to electronic movements can be neglected to a first approximation, and the probability that they will be in the volume $d\tau_a = dX_1 \ldots dZ_p$ can be considered to be independent of the probability that the electrons will be in the volume $d\tau_e = dx_1 \ldots dz_n$. The function ψ, therefore, may be taken as the product of a function ψ_a of the nuclear coordinates and a function ψ_e of the electronic coordinates.

According to this hypothesis, the nuclei behave as mass points in a stationary field. Thus, without knowledge of the movement of electrons, the geometrical configuration of the nuclei and the field of force maintaining them in stable equilibrium can be determined. The problem now is the well-known one of the movements of a system of point masses about their positions of equilibrium (Whittaker). If the deformations are small enough only the terms of second degree in the potential energy of the system need be retained, and classical mechanics completely solves this simplified problem.

Under these conditions, when a molecule (undergoing thermal agitation)

is given a weak impulse, the nuclei acquire kinetic energy and, from a state of stable equilibrium, the whole system performs a complex movemen in relation to axes dependent on the molecule, which can be taken as a superposing of 3*s*–6 sinusoidal and independent vibrations, in the course of each of which all the nuclei vibrate in phase on rectilinear trajectories. The fundamental vibrations, which are superposed to give the general motion, have arbitrary phases and amplitudes; each is defined by a normal coordinate $q_i = Q_i \cos(\omega_i t - \phi_i)$, on which the displacement of the nuclei depends linearly.

The spectral analysis of the light absorbed or scattered by the molecules of a fluid is but a harmonic analysis of the movements of these molecules. However, all fundamental vibrations do not interact with the incident light wave and all molecular frequencies do not appear in the spectrum; i.e., we have the selection rules relating to absorption spectra (Brester, 1923) and to scattering spectra (Placzek, 1931; Cabannes, 1932).

(1) In order to give an absorption band in the infra-red spectrum, a fundamental vibration of frequency ω_k must give rise to a variable electric moment μ; the molecule can then absorb radiation of frequency ω_k by resonance, and it is said to be active in absorption.

(2) The scattered radiation is produced by variation of the electric moment, $m = [\alpha]\,\epsilon_0 \sin \Omega t$, induced in the molecule by the electric field $\epsilon_0 \sin \Omega t$ of the incident light wave. Consider a fundamental vibration of frequency ω_k, defined by the normal coordinates $q_k = Q_k \cos(\omega_k t - \phi_k)$; in order that the Raman lines with a frequency $\Omega \pm \omega_k$ may appear in the spectrum, it is necessary and sufficient for the vibration considered to cause a change in the tensor $[\alpha]$ or, in other words, that the derived tensor $\dfrac{\partial[\alpha]}{\partial q_k}$ should not be zero.

A polarization law is added to this selection rule so that the following three cases can be distinguished :

$\dfrac{\partial[\alpha]}{\partial q_k} = 0$	inactive vibration	forbidden line
$\dfrac{\partial[\alpha]}{\partial q_k} \neq 0$	$\dfrac{\partial(A+B+C)}{\partial q_k} = 0$	depolarized line $\rho = 6/7$
	$\dfrac{\partial(A+B+C)}{\partial q_k} \neq 0$	polarized line $\rho < 6/7$

$A+B+C$ represents the sum of the main components of polarizability, A, B, C, of the molecule, i.e. the spur of the matrix $[\alpha]$. The depolarization factor ρ measures the relation between the intensities of the smallest and largest incoherent rectangular vibrations which constitute the radiation emitted at right angles to the incident beam of ordinary light.

A vibration could therefore be inactive for scattering (I) or give a Raman line either depolarized (D) or polarized (P); it will be inactive (I) or active (A) for absorption. In this way six kinds of vibration, designated by II, DI, PI; IA, DA, PA, can be distinguished by their action on light. The interest of these rules and this classification is in the part played by them in research on molecular symmetry.

(ii) *Determination of molecular symmetry*

The most striking fact about spectroscopic analysis of nuclear vibrations is that the symmetry of the molecule at rest is clearly shown without any hypothesis on the nature of the connecting forces being formulated. Now all the physical and chemical properties of a molecule are dependent on its symmetry. From this arises the importance of the fruitful application of the Raman effect to chemistry. The symmetry alluded to is that of the arrangement of the nuclei, assumed to be stationary in their positions of equilibrium; it shows the symmetry of the force field; it is defined by a number of symmetry elements (axes, planes, centres), and the fundamental vibrations are classified into different types according to the symmetry elements of the molecule at rest, which are preserved in movement; there are found, in this way, totally symmetrical vibrations and others which lose elements of symmetry in the process.

If the symmetry element, which disappears during the vibration, is a centre, plane or binary axis, the vibration must be antisymmetrical with relation to that element; in other words, at the end of the half-period the nuclei of the deformed molecule are in a position symmetrical to that at the beginning. Moreover, when the axis of the molecule is of an order higher than 2, certain natural frequencies merge in pairs, giving rise to degenerate vibrations in which an atom can describe any ellipse in the plane fixed by the mechanical conditions. We shall have, therefore, to consider several types of vibration.

The effect of the vibrations on the electric moment μ and the tensor $[\alpha]$, which is different from one type to another, conforms to the selection rules and polarization characteristic of each of them. In that connected with the scattering of light by a fluid, in which the molecules are not preferentially orientated:

(i) Only completely symmetrical vibrations give a polarized Raman line $\rho<6/7$.

(ii) The vibrations which are antisymmetrical in relation to a binary axis or a plane give depolarized Raman lines ($\rho=6/7$); the same rule applies to degenerate vibrations.

(iii) Vibrations which are antisymmetrical in relation to a centre are inactive.

(iv) Vibrations which are symmetrical in relation to a multiple axis and antisymmetrical in relation to symmetry elements of a different class are inactive. The same rule applies to vibrations which are degenerate owing to the presence of a multiple axis, and which are symmetrical to a plane normal to this axis.

For absorption:

(i) If vibrations which are symmetrical in relation to an axis (or plane) cause a variable electric moment μ, this vector is situated on the axis (or perpendicular to the plane).

(ii) If vibrations, which are antisymmetrical in relation to an axis (or plane), cause a moment μ, the vector is perpendicular to the axis (or in the plane).

(iii) Vibrations which are symmetrical to a centre are inactive.

It is generally easy to determine the different types of fundamental vibration of a symmetrical molecule. In order to enumerate the vibrations of each type

it is sufficient to count the degrees of freedom compatible with the conditions of symmetry characterizing the type of movement considered and to subtract the translations and rotations of the same sort. There are, for example, three completely symmetrical vibrations of the ethane molecule C_2H_6, its being regarded as having either the D_{3h} or D_{3d} symmetry; indeed, when the displacement of a H nucleus in the plane through the three-fold axis (two variables) and that of a C nucleus on the axis (one variable) have been defined, the displacement of all the other nuclei is determined by the conditions of symmetry.

The preceding rules allow a systematic determination of the symmetry of a molecule, of which the Raman and infra-red spectra have been observed. Observed spectra can be compared with those predicted for different conceivable and reasonable molecular models (Kohlrausch, 1938). Other physical methods (measurement of dielectric constant, electrical birefringence, etc.) are no more than means of control. Most of the time neither the rotation structure of the infra-red bands nor even the outline of this structure are known, because the substance has not been studied in the vapour phase; the information obtained from the orientation of the variable electric moment is lacking. One is most often restricted to measuring the depolarization factor of the Raman lines, which is sensibly the same in the liquid and vapour states if intermolecular actions are weak.

(iii) *Application of group theory*

The formulae of the group theory give, very rapidly, the different types of vibration, their symmetry, their action on light and the number of frequencies of each type. This theory was postulated by Wigner (1930, 31), applied to an actual problem by Placzek (1934) and systematically expounded by Rosenthal and Murphy (1936).

The covering operations of a molecule to a state of rest constitute a group G. To each element R of this group there corresponds a matrix Γ_R which is constructed in the following way. Let $V = \Sigma \frac{1}{2}\omega^2(q_1^2 + q_2^2 \ldots + q_k^2)$ be the expression for the potential energy of the deformed molecule as a function of the fundamental frequencies ω and the normal coordinates q. A covering operation R (which can always be considered to be a rotation of the first or second type) converts each series of coordinates $q_1, q_2, \ldots . q_k$ into another series of orthogonal co-ordinates. Let γ_R be the matrix defining this partial transformation. The matrix which transforms the whole set of the normal coordinates is written

$$\Gamma_R = \begin{pmatrix} \gamma_R & & \\ & \gamma_R' & \\ & & \gamma_R'' \end{pmatrix}; \qquad \ldots\ldots(1)$$

each of the matrices $\gamma_R, \gamma_R', \gamma_R'' \ldots .$ corresponds to one of the distinct frequencies $\omega, \omega', \omega'', \ldots .$, and the shaded part only contains zeros.

If all the operations R of the group G are considered, the set of matrices γ_R relative to the frequency ω is, by definition, an irreducible representation of the group; therefore there are as many irreducible representations as distinct frequencies ω. However, several series of normal coordinates of different frequencies can be transformed in the same way by all the operations of the group G. The vibrations defined in this way are of the same type; therefore there are as many types of vibration of the molecule as there are distinct irreducible representations in the group G. It is shown that this number is equal to the number of classes of the group.*

The character $X_i(R)$ of the operation R in the irreducible representation γ_R^i is the spur (i.e. the sum of the diagonal terms) of the matrix. This number characterizes the way in which the operation R transforms the vibrations of type i. In the case of a non-degenerate vibration, the normal coordinate q is changed to $\pm q$, and the character is equal to ± 1, depending on whether the vibration is symmetrical or antisymmetrical with reference to the symmetry element of the molecule at rest. In the case of a doubly degenerate vibration, two normal coordinates q_1 and q_2 are interposed; the matrix has two ranges and the character can have the values 0, ± 1, ± 2, depending on the transformation which the system of coordinates (q_1, q_2) has undergone during the operation R. We will show, with two examples, how the symmetrical properties of each type i of the vibrations and their characters $X_i(R)$ are obtained. These two examples will be used later on.

The group $D_{3h}(\bar{6}\,2m)$† includes the following operations: the identical operation E; one reflection Σ_h on the symmetry plane normal to the principal axis; two rotations S_6 about the same axis considered as an axis $(\bar{6})$; two rotations C_3 about the same axis considered as an axis (3); one rotation C_x about each of three axes (2) normal to the principal axis; one reflexion Σ_v on each of three symmetry planes (m) passing through the principal axis. These twelve operations are divided into six classes: E, Σ_h, S_6, C_3, C_x, Σ_v; therefore there are six types of vibration. The group considered being one of eleven primary groups (which could not be obtained as direct products of simpler groups), it is necessary to determine once for all the symmetry properties of each of the six predicted types, from which the $X_i(R)$ characters are derived. These characters are shown in columns 2 to 7 of table 2.

The group $D_{3d}(\bar{3}\,2/m)$ is the direct product of the primary groups D_3 and I, of which the elements are E, $2C_3$, $3C_x$, and E, I (the designation I indicates inversion with respect to a centre). There are six elements of the direct product:

$$E \times E = E;\ E \times I = I;\ 2(C_3 \times E) = 2C_3;\ 2(C_3 \times I) = 2S_3;$$
$$3(C_x \times E) = 3C_x;\ 3(C_x \times I) = 3\Sigma_x;$$

* If the relation between three elements A, B, X of the group is $B = XAX^{-1}$, it is said that the elements A and B are conjugate. A set of conjugated elements forms a class of the group. For example, corresponding to the covering group D_{3h}, from binary normal axes to a multiple axis, is a class of three conjugated operations. Similarly, rotations of 120° and 240° about the multiple axis are conjugated owing to the presence of binary axes, and form another class of the group.

† The notation adopted by the Zürich Conference for covering groups and their repetitive units is indicated in parentheses (Mauguin, 1931).

there are, therefore, six types of vibration.* Each type of vibration of the group D_3 is divided into two types of vibration of the group D_{3d}, one symmetrical and one antisymmetrical to the centre. The characters $X_i(R)$ of the group D_{3d} can be directly calculated, or they can be obtained by starting with the characters of the primary group D_3 and multiplying by $+1$ or -1, according as the element considered of group D_3 has been multiplied by E or I. Those characters are shown in columns 2 to 7 of table 3.

Table 2

D_{3h} ($\bar{6}$ 2 m) $h=12$		E	Σ_h	$2S_6$	$2C_3$	$3C_2$	$3\Sigma_y$	a	Scattering	Absorption	Translations	Rotations	Vibrations
	A'_1	1	1	1	1	1	1	3	A	I	0	0	3
	A_1''	1	−1	−1	1	1	−1	1	I	I	0	0	1
	A_2'	1	1	1	1	−1	−1	1	I	I	0	1	0
	A_2''	1	−1	−1	1	−1	1	3	I	A	1	0	2
C_3X_6	E'	2	2	−1	−1	0	0	4	A	A	1	0	3
α form	E''	2	−2	1	−1	0	0	4	A	I	0	1	3
$\cos\phi$		1	1	$-\frac{1}{2}$	$-\frac{1}{2}$	−1	1						
$\pm1+2\cos\phi$		3	1	−2	0	−1	1						
u		8	0	0	2	0	4						
$X(R)$		24	0	0	0	0	12						
$\cos\phi(\pm1+2\cos\phi)$		3	1	0	0	1	1						

Table 3

D_{3d} ($\bar{3}$ 2/m) $h=12$		E	I	$2C_3$	$2S_3$	$3C_2$	$3\Sigma_y$	a	Scattering	Absorption	Translations	Rotations	Vibrations
	A_{1g}	1	1	1	1	1	1	3	A	I	0	0	3
	A_{1u}	1	−1	1	−1	1	−1	1	I	I	0	0	1
	A_{2g}	1	1	1	1	−1	−1	1	I	I	0	1	0
	A_{2u}	1	−1	1	−1	−1	1	3	I	A	1	0	2
C_3X_6	E_g	2	2	−1	−1	0	0	4	A	I	0	1	3
β form	E_u	2	−2	1	−1	0	0	4	I	A	1	0	3
$\cos\phi$		1	−1	$-\frac{1}{2}$	$-\frac{1}{2}$	−1	1						
$\pm1+2\cos\phi$		3	−3	0	0	−1	1						
u		8	0	2	0	0	4						
$X(R)$		24	0	0	0	0	12						
$\cos\phi(\pm1+2\cos\phi)$		3	3	0	0	1	1						

* Placzek's notation has been used for the various types of vibration. A, a vibration symmetrical to a ternary axis; E, a degenerate oscillation. A_1 (or A_2), vibrations symmetrical (or antisymmetrical) to binary axes normal to the ternary axis. A' (or A''), vibrations symmetrical (or antisymmetrical) to a symmetry plane normal to the ternary axis. A_g (or A_u), vibrations symmetrical (or antisymmetrical) to a centre.

The next problem is to determine the number of vibrations of each type. The matrix Γ_R has been obtained by allowing the operator R to act upon the normal coordinates; a different matrix would be obtained starting from Cartesian coordinates. However, the sum of the principal diagonal terms would remain constant. This important number $X(R)$ is the character of the operator R. From equation (1):

$$X(R) = a_1 X_1(R) + a_2 X_2(R) + \ldots = \Sigma a_i X_i(R). \qquad \ldots\ldots(2)$$

$X_1(R), X_2(R), \ldots$ are the characters of the elements of the range R in the distinct irreducible representations of group G (see, for example, tables 2 and 3); $a_1, a_2 \ldots$ represent the number of distinct frequencies for each type of vibration.

On the other hand, if R operates on the Cartesian coordinates of the nuclear displacements, a numerically calculable expression of the character $X(R)$ is obtained:

$$X(R) = u_R(\pm 1 + 2\cos\phi_R). \qquad \ldots\ldots(3)$$

ϕ_R is the angle of the rotation R with a + or − sign, depending on whether the operation is of the first or second kind; u_R is the number of nuclei not displaced by this operation.

Finally, it is shown that the characters $X(R)$ form an orthogonal system; it follows that

$$a_i = \frac{1}{h}\Sigma_R\, X(R).X_i(R), \qquad \ldots\ldots(4)$$

where the summation extends to the h elements of the group G. The characters $X(R)$, which were calculated from equation (3), are shown in the ninth line of tables 2 and 3; values of a_i (from equation (4)), are shown in the eighth column.

Generally the displacement of the nuclei causes a variable electric moment μ and a variation in $[\alpha]$, the tensor. Let us assume first that at each instant the vector μ and the tensor $[\alpha]$ remain in the symmetry of the nuclear configuration, and secondly that the three components of the vector and the six coefficients of the tensor are linearly related to the coordinates defining the nuclear displacement. According to this hypothesis three matrices can be allotted to each operation R of the group G:

(i) the matrix Γ_R, which changes the nuclear coordinates;
(ii) the matrix Γ'_R, which changes the vector μ;
(iii) the matrix Γ''_R, which changes the tensor $[\alpha]$.

Each of these matrices is a separate representation of the operation R. The character of the first is known (3). The characters of the second and third are:

$$X'(R) = \pm 1 + 2\cos\phi_R, \qquad \ldots\ldots(5)$$

$$X''(R) = 2\cos\phi_R(\pm 1 + 2\cos\phi_R). \qquad \ldots\ldots(6)$$

The number of times an irreducible representation $\Gamma^{(0)}$ of the group G is contained in the representation $\Gamma^{(v)}$ can be calculated; it is found to be

$$a_i' = \Sigma_R(\pm 1 + 2\cos\phi_R)X_i(R). \qquad \ldots\ldots(7)$$

If this number is zero, the vibrations of type i do not cause an electric moment and are inactive in absorption.

In the same way, the number of times the irreducible representation $\Gamma^{(0)}$ of group G is contained in the representation Γ'' can be calculated; it is found to be

$$a_i'' = \Sigma_R 2\cos\phi_R(\pm 1 + 2\cos\phi_R)X_i(R). \qquad \ldots\ldots(8)$$

If this number is zero, the vibrations of type i do not vary the tensor $[\alpha]$ and are inactive in scattering. An application of these selection rules is shown in columns 9 and 10 of tables 2 and 3.

These tables show the difficulty of attempting to decide between the α or β form of a molecule like ethane or its halogen derivative C_2X_6. The only property which will be able to differentiate between them is that the former has three degenerate frequencies occurring in both the Raman and infra-red spectra.

§ 5. DETERMINATION OF STRUCTURES

Spectral analysis is the most usual and undoubtedly the most efficient of all the methods for determining the symmetry of molecules containing a sufficiently small number of atoms. The different types of vibration, and the number of vibrations of each type, are sought for each possible structure. The number of inactive and active frequencies in absorption and scattering are predicted; so are the frequencies which appear in both spectra, and the polarized and depolarized lines in the scattering spectrum. It only remains to determine to which symmetry group the observed spectra correspond.

When the molecule contains a small enough number of atoms, the arrangement of nuclei is determined without doubt by the architectural symmetry. In this way, among molecules with a ternary axis, some (ClO_3^-, NH_3, PCl_3) have C_{3v} symmetry and form a triangular pyramid because the two non-degenerate frequencies are active in both absorption and scattering; others (CO_3^{--}, NO_3^-, BCl_3) have D_{3h} symmetry and are planar since of the two non-degenerate frequencies, one is active in absorption, the other in scattering.

However, there are great difficulties in most cases. Optical spectra do not reproduce the spectrum of fundamental frequencies exactly. Lines due to slightly active vibrations do not appear on the negatives; the anharmonicity of intramolecular movements complicates infra-red spectra and causes additional Raman lines; mere enumeration of the observed frequencies is of no value. On the other hand, measurement of the depolarization factor is not always possible. The more complicated the molecule the less symmetry it has; it follows that the rules in the preceding paragraph are no longer applicable to the most difficult problems; if the molecule has no element of symmetry it is obvious that it will have only one type of vibration.

An attempt is made to classify vibrations according to the atoms which participate in the movement and the direction of their trajectories. The characteristic frequencies and small displacements of nuclei can be predicted from simple hypotheses about the nature of the forces. In particular, valency forces, in accordance with chemists' ideas about molecular structure, give a good

approximation in many cases: the observed frequencies are found in the immediate vicinity of calculated frequencies. These calculations, however, are not always necessary; specialists in the Raman effect can discover the origin of certain lines by comparing many spectra. As more compounds are observed, spectral analysis becomes easier and more fruitful.

(i) *Empirical rules*

The allocation of frequencies has been made much more simple by the following more or less empirical rules:

(i) The forces due to deformation of valency-bond angles are usually weaker than those due to the displacement of atoms along these bonds. In these circumstances deformation vibrations and valency vibrations can be distinguished according to whether the former or the latter forces are in operation; also, deformation vibrations have a lower frequency than valency vibrations. For hydrogen these frequencies are of the order 1400 and 3000 cm^{-1} respectively.

(ii) Certain radicals retain approximately the same characteristic frequencies in different molecules. For this they must be only weakly bound to the rest of the molecule. More often the vibrations of two radicals in the same molecule are not coupled if they are of different types as to molecular symmetry, and coupling between two vibrations of the same type is weak when their frequencies are widely separated. This is the case for the vibrations of hydrogen along the valency bond, and they can always be treated separately. This separation of high frequencies makes it possible to treat the units CH_3, CH_2, NH_2, NH, OH, ... as single atoms for the study of a chain in organic chemistry. The existence of symmetry elements made it possible to break down the frequency equation rigorously into as many algebraic equations of lower degree as there are types of vibration. In its turn each of these partial equations can be broken down, as a first approximation, into equations of high and low frequencies (Wilson, 1941).

(ii) *Isotopic displacements*

As has just been shown, the separation between two frequencies is produced sometimes by a difference in the forces acting, sometimes by a difference in the moving masses. If the forces are unknown, the masses constitute the data of the problem, and one is led to seek how the frequencies of the molecules are modified when the masses of their constituent atoms are varied systematically. The method becomes especially efficient and certain when the hydrogen of a hydrogenated molecule can be replaced by deuterium. In this case the change of frequency is sufficient for easy observation of the isotopic effect and small enough for certain recognition of the modified frequencies. On the other hand, the field force is unchanged, making a quantitative verification possible. The product rule (Redlich, 1935; Teller, 1935; Wilson, 1941) has been very useful in the classification of frequencies.

The constant term in the frequency equation can be calculated in two ways:

it is equal to the product of the square of the frequencies; it is also equal to the quotient of the determinants formed with the matrices A and B from the quadratic expressions, in Cartesian coordinates, for the potential and kinetic energies of the deformed molecule:

$$\omega_1^2 \times \omega_2^2 \times \ldots . \omega_n^2 = \frac{|A|}{|B|},$$

where $|B|$ is equal to the triple product of the masses $m_1 \times m_2 \times \ldots . m_s$ of the nuclei and $|A|$ is dependent only on the field of force. Then, by substituting isotopes for certain atoms, the following relation is obtained between the characteristic frequencies of the molecules:

$$\frac{\omega_1 \times \omega_2 \times \ldots . \omega_n}{\omega_1' \times \omega_2' \times \ldots . \omega_n'} = \sqrt{\frac{m_1' \times m_2' \times \ldots . m_s'}{m_1 \times m_2 \times \ldots . m_s}}.$$

The ratio of the products of the fundamental frequencies of the two molecules is inversely proportional to the square root of the ratio of the products of the atomic masses. When the frequency equation is broken down by symmetry (and if the symmetry is preserved in isotopic substitution) the rule is separately applicable to each type of vibration. If substitution only takes place on some of the nuclei which are symmetrically placed in the original molecule, the decrease in symmetry of the molecule doubles the degenerate lines and causes frequencies, originally inactive, to appear in optical spectra. By applying these methods to benzene and deuterobenzene, Angus and his colleagues (1936) were able to classify the frequencies with certainty. Sutherland, Lee and Wu (1939), and Cameron, Sears and Nielsen (1939) used these methods to clasify the infra-red bands of the molecules PH_3 and PD_3, AsH_3 and AsD_3, and SeH_2 and SeD_2, although the analysis of the rotational structure was incomplete; Gallaway and Barker (1942) have compared the spectra of C_2H_4 and C_2D_4; Wilson (1943) those of nitro- and trideuteronitromethanes; Redlich and Nielsen (1943) those of the acids NO_3H and NO_3D.

Except in the case of hydrogen and deuterium, isotopic effects are too feeble to be commonly used. However, the most recent spectographs, which have high luminosity and good dispersion, give new possibilities. Wagner resumed the study of simple molecules in the liquid state (SO_2, COS, BrCN, BCl_3 and BBr_3); he found very weak lines of various origins (isotopic doublings, frequencies of the second order, lines of unknown origin); from isotopic doubling he could draw certain conclusions. Thus, in boron trichloride the intervals between the components of the strong, polarized, Raman line [472·1 ($B^{35}Cl_3$); 468·0 ($B^{35}Cl_2{}^{37}Cl$); 464·3 ($B^{37}Cl_3$)] show that the boron atom does not take appreciable part in the corresponding symmetrical vibration, and hence the molecule is planar. This conclusion is naturally extended to the BBr_3 molecule, although the isotopic intervals here cannot be observed.

(iii) *Frequency changes*

An easier method of classification is to observe the frequency changes when the mass of certain atoms is progressively increased or diminished by substitution, while the rest of the molecule is unchanged. Daure carried out this type of research, immediately after the discovery of the Raman effect, on the trihalides XY_3 (X = P, As, Sb; Y = Cl, Br) and the tetrahalides XY_4 (X = C, Si,

Ti, Sn; Y = Cl, Br). In 1932 Cabannes and Rousset were able to show the symmetry of the molecules CX_4 (T_d), CX_3Y (C_{3v}), CX_2Y_2 (C_{2v}) by spectral analysis of the halogenated derivatives of methanes. Since that time these molecules have been the subject of much research. Kohlrausch and Wagner (1939) studied the series CH_4, CH_3X, CH_2X_2, CHX_3, CX_4; $C(CH_3)_4$, $C(CH_3)_3X$, $C(CH_3)_2X_2$, $C(CH_3)X_3$, CX_4; CH_2XY(X, Y = CH_3, Cl, Br, I). Glockler and Leader (1940) investigated the series $CHBr_3$, $CHBr_2Cl$, $CHBrCl_2$, $CHCl_3$; $CHFBr_2$, CHFBrCl, $CHFCl_2$; CFH_3, CFH_2Cl, $CFHCl_2$, $CFCl_3$. Glockler and Edgell (1941) studied the series CHX_3(X = H, D, F, Cl, Br, I) and CH_4, CH_3F, CH_2F_2, CHF_3, CF_4. Delwaulle and François (1941–45) studied the series CCl_4, CCl_3Br, CCl_2Br_2, $CClBr_3$, CBr_4; $CFBr_3$, CF_2Br_2, $CFBr_2Cl$, $CFBrCl_2$, $CFCl_3$; $CHFCl_2$, CHFClBr, $CHFBr_2$; CH_2Br_2, CH_2BrCl, CH_2BrF. Most recently, the latter authors have investigated the molecules $POCl_3$, $POCl_2Br$, $POClBr_2$, $POBr_3$, and this preliminary study has enabled them to discover the reason for errors made by previous authors about the halides AX_3, AX_2Y, AXY_2, AY_3(A = C, Si, Ti, Sn; X = Cl, Y = Br). The derivatives of carbon and silicon are very similar, whereas titanium is more like tin; in particular, titanium chlorobromides (like stannic chlorobromides already observed by Trumpy (1931), cannot be isolated by chemical means and can be detected in mixtures only by spectral analysis.

The displacement of each frequency is easily followed when one atom is substituted for another of the same valency in a molecule the symmetry of which remains unchanged; if there is irregularity in the "frequency–atomic mass" curve, the substitution has changed not only the moving mass, but also the field of force, i.e. the electronic configuration.

When the symmetry is changed, the frequency changes become more difficult to follow; nearly every individual case has to be examined. One of the atoms may be much lighter than the others, in which case the frequency of vibration along the bond is unchanged. Thus the C—H frequency retains the value 3021 throughout the series $CHBr_3$, $CHBr_2Cl$, $CHBrCl_2$, $CHCl_3$; similarly the C—H and C—F frequencies remain at 3017 and 1065 in the series $CHFBr_2$, CHFBrCl, $CHFCl_2$. However, generally all the atoms are displaced during each fundamental vibration and the problem must be approached in a general way.

Let us follow the frequency change in the molecules

$$CCl_4 \longrightarrow CCl_3Br \longrightarrow CCl_2Br_2 \longrightarrow CClBr_3 \longrightarrow CBr_4.$$

Table 4 and its adjoining notes show how correspondences are established; however, there are anomalies and uncertainties. Thus, all four of the frequencies 716–778 of CCl_3Br and 680–742 of $CClBr_3$ appear to be polarized ($\rho = 0{\cdot}70$); the most intense lines, 716 and 680, could be rather arbitrarily assigned to the completely symmetrical vibration A_1; but, if $CClBr_3$ is compared with the molecules $SiClBr_3$ and $SnClBr_3$, in which one polarized and one unpolarized line are observed, it can be seen that the polarized line (i.e. the line A_1) is that with the higher frequency and lower intensity:

$$CClBr_3 \begin{cases} 680\ (\rho = 0{\cdot}70) \searrow \\ 742\ (\rho = 0{\cdot}70) \nearrow \end{cases} 673 \qquad SiClBr_3 \begin{cases} 498\ D \searrow \\ 579\ P \nearrow \end{cases} 487 \qquad SnClBr_3 \begin{cases} 283\ D \searrow \\ 388\ P \nearrow \end{cases} 279$$

Table 4

Molecules	CCl_4	CCl_3Br	CCl_2Br_2	$CClBr_3$	CBr_4
Symmetry group	T_d	C_{3v}	C_{2v}	C_{3v}	T_d
Symmetry elements	"	$[C_3, \Sigma]$	$[\Sigma, C_2, \Sigma']$	$[\Sigma', C_3']$	"
E	217 D → E	192 D → A_1	154	E	141 D → E	123 D
		→ A_2	174		
			B_1	230 D		
F_2	313 D → A_1	248 D → A_1	241 P → A_1	214 P → F_2	182 D
	→ E	295 D → B_2	262 D → E	270 → F_2	
A_1	460 P → A_1	422 P → A_1	380 P → A_1	328 P → A_1	268 P
			B_1	683 D → E	680 → F_2	673 D
F_2	789 D → A_1	716 → A_1	733 P → A_1	742	
	→ E	778 → B_2	770 D		

C_3 . . . axis through A
Σ . . . plane ABH
C_2 . . . axis HK
Σ' . . . plane CDK
C_2' . . . axis through C

For the notation of the types of oscillation A_1, A_2, B_1, B_2 E, F_2, see Kohlrausch, 1938, p. 41.

The letters P or D indicate that the Raman line is polarized or unpolarized.

The vibration E of the group T_d, symmetrical to C_3, is doubled in a vibration A_1, symmetrical to Σ_y (PA) and a vibration A_2, antisymmetrical to Σ_y (DI). As there is only one frequency of the type A_2 in the molecule CCl_2Br_2, all the A_1 frequencies of the group C_{2v} will give A_1 (PA) frequencies of the group C_{3v}. Then the E frequencies of the C_{3v} group become doublets, B_1 and B_2 (DA).

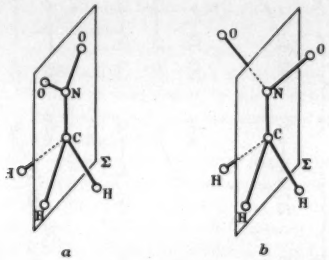
It remains to explain the following observed anomalies: the doubling of the CCl_4 line 762–790 and the polarization $\rho = 0{\cdot}79$ of its components; the polarization $\rho = 0{\cdot}65$ of the CCl_2Br_2 lines 154–174; the polarization $\rho = 0{\cdot}70$ of the CCl_3Br lines 716–778 and the $CClBr_3$ lines 680–742.

When the depolarization factor cannot be measured, the correspondence is not certain. An approximate calculation of the frequencies is made starting from a simple hypothesis on the nature of the forces; in this way a frequency equation is obtained in which the atomic masses are variable parameters. In the case of very symmetrical molecules the calculation of the frequencies of each type is relatively easy and their variation is followed as a function of one of the masses.

(iv) *Some recent determinations of structures*

(*a*) CH_3NO_2

An example of these systematic studies is found in the writings of Chédin (1939), Mathieu and Massignon (1941), Lecomte and Mathieu (1942) on nitric acid, nitro-paraffins and nitric esters. In the $CH_3.NO_2$ molecule vibrations



Nitro-methane molecule

Figure 2.

can be distinguished, during which one of the two radicals only, either CH_3 or NO_2, is deformed; during others these radicals are displaced with respect to each other without any appreciable deformation. These are therefore classified into internal frequencies, characteristic of the radicals CH_3 and NO_2 respectively, and external frequencies, characteristic of their linkage. When isolated, the CH_3 radical (C_{3v}) gives two polarized Raman lines at 1300 and 2950 and two depolarized lines at 1450 and 3000; the NO_2 radical (C_{2v}) gives two polarized lines at 600 and 1350 and one depolarized line at 1550. As has been stated already, these radicals retain some individuality in the sense that their internal frequencies are slightly modified in different molecules, but the disymmetry of a molecule affects, to some extent, that of its constituent radicals. Thus, in the molecule $(CH_3)CHBr(NO_2)$, which has no symmetry element, the internal frequencies of the CH_3 radical are no longer degenerate and all the Raman lines are polarized.

From the analysis of the Raman lines of the molecule H_3CNO_2 it was deduced that it only exists in one form: the two parts of the molecule cannot rotate freely with respect to each other about the C—N axis, and the Raman line of low frequency 483 is due to torsion about this axis.

Apart from the vibrations of the hydrogen nuclei, the six remaining vibrations are those of a system of four masses CNO_2 situated in the same plane P (figure 2). The internal vibrations of the CH_3 radical are no longer degenerate; three Raman lines are observed in the regions 2679–2967–3056. Since depolarized lines are found, the molecule must have a symmetry element Σ, which can only be the plane P or a symmetry plane perpendicular to this plane. From the fact that the antisymmetrical vibration 1562 of the NO_2 radical remains depolarized in the molecule it is concluded that it is antisymmetrical with respect to Σ. Therefore, the plane P does not coincide with Σ but is perpendicular to it (figure 2*b*).

(*b*) $C(NO_2)_4$

Tetranitromethane gives a good example of the determination of molecular symmetry by spectral analysis. Mathieu found, at the maximum, six polarized Raman lines all inactive in absorption, indicating a high degree of symmetry. To each of the peaks of the carbon tetrahedron a NO_2 must be attached in such a way that the CNO_2 arrangement is planar. Therefore the problem is to orient the ONO triangle in the correct azimuth around the C—N bond. Through this axis a symmetry plane Σ_v of the tetrahedron passes and there are three possible hypotheses: the NO_2 radical is in the plane Σ_v or perpendicular to this plane (symmetry S_{4h}), or it may make a certain angle between 0° and 90° (symmetry S_4) with the plane Σ_v.

In the first two models there are five types of vibration, but the number of vibrations of each of the non-degenerate types varies according to whether the two oxygen atoms are on the symmetry plane Σ_v or symmetrically placed with respect to Σ_v. Table 5 gives the different types, their optical properties and the number of frequencies for each of them.

Table 5

	S_4	C_2	Σ_v		Plan NO_2		
					Coincident with Σ_v	Perpendicular to Σ_v	Inclined to Σ_v
A_1	sym.	sym.	sym.	PI	6	5	8 PI
A_2	,,	,	antisym.	II	2	3	
B_1	antisym.	,,	,,	DI	3	4	9 DI
B_2	,,	,,	sym.	DA	6	5	
E	deg.	antisym.	deg.	DA	8	8	8 DA

In the spectra the frequencies ω and ω' of the symmetrical and antisymmetrical vibrations along the N—O bonds are recognized; but, owing to linkages in the molecule between the movements of the NO_2 radicals, each of the internal vibrations of the isolated radical gives rise to three molecular vibrations of types *A*, *B*, *E* respectively.

If the vibration of the NO_2 radical is symmetrical with respect to the plane Σ_v which passes through the N nucleus, it will give rise to a movement of the A_1 type in the molecule, producing a polarized Raman line which is inactive in the infra-red. Obviously it is the case for the vibration ω which gives the lines 1348 ($\rho = 0{\cdot}07$) and 1273 D in the Raman spectra of the molecule and the line 1245 in the infra-red.

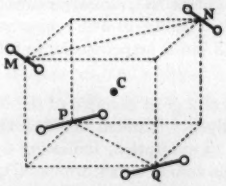


Figure 3.

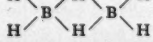
The four nitrogen atoms equidistant from the carbon atom on the lines CM, CN, CP CQ have not been shown.

If the vibration of the NO_2 radical is antisymmetrical with respect to the plane Σ_v of the molecule, it does not give a polarized Raman line. The vibration ω' is found to give the Raman lines 1648 D and 1614 D and the infra-red line 1587.

During the vibration ω' the NO_2 nuclei are displaced in their plane P. Since the vibration is antisymmetrical with respect to the symmetry plane Σ_v, Kohlrausch and Wittek (1941) have concluded that the two planes P and Σ_v are rectangular (figure 3).

(*c*) B_2H_6 and Al_2Cl_6

Wagner (1943) has formulated the more difficult problem of the structure of diborane. There are two possible arrangements:

(i) and (ii) 

The first has C_{2h} or C_{2v} symmetry; the second, D_{2h} symmetry with two rectangular symmetry planes, one of which contains the H nuclei of the closed chain, the other the external H nuclei.

By using reasonable values of the valency forces the frequencies for each of the two models can be predicted by calculation. The symmetrical frequencies found in this way for model (ii) are :

A_{1g} 593–616 1079–1094·5 2131–2138·5 2644–2646;

the first number of each doublet refers to the molecule B^{11}–B^{11} and the second to the isotopic molecule B^{10}–B^{10}. In the Raman spectrum two intense lines

are observed, one at 2101, the other at 2523, the presence of which is hard to explain by model (i), where all the H nuclei play the same part.

However, isotopic displacements contribute nothing to the solution of this problem, and the assignment of infra-red bands presents difficulties.* In order to confirm the correctness of model (ii) the infra-red spectrum must be re-investigated to determine the contour of the bands, and it is necessary to measure the polarization of the Raman lines. Kohlrausch and Wagner (1942) have reported analogous difficulties in the determination of the structure of the dimeric molecules Al_2Cl_6 and $Al_2(CH_3)_6$, to which they had already attributed D_{2h} symmetry, although recognizing the difficulty of explaining the union of two Al atoms by two intermediate CH_3 groups.

§ 6. ROTATIONAL ISOMERISM

It is impossible to characterize cis- and trans- isomers chemically unless the bonds fix the atoms so firmly that the configuration does not change during the reaction. This is the case with dibromo-ethylene BrCH=CHBr, in which the carbon atoms are joined by a double bond. The isomers are recognized by the acids formed by reacting them first with potassium cyanide and then with a strong base in the presence of water. Only in one of the two acids do the two-carboxyl groups react with each other to give an anhydride, so it is assumed that they are adjacent and, therefore, that it is the cis-form. When the carbon atoms are singly linked, as in dibromo-ethane $BrCH_2$—CHBr, it is impossible to distinguish chemically between the two isomers, so it has been assumed that there is free rotation of both halves of the molecule about the C—C axis with no preferred orientation. Ideas started developing in 1937 when Sidgwick suggested that two singly linked carbon atoms revolve in relation to each other (each carrying the atoms bound to it) to take up the most stable configuration, with the result that only one configuration is possible.

If the two halves of the molecule are imagined to rotate about the C—C axis with respect to each other, the potential energy is found to be stationary for certain orientations *A* (cis) and *D* (trans) (figure 4). In fact it is possible to conceive that the inertaction of the halogen atoms is not sufficient to alter appreciably the potential barriers that fix the configuration of the ethane molecule. It is then possible for each of the six orientations, *A*, *B*, *C*, *D*, *E*, *F*, to correspond to a minimum of energy, in which case there are several stable forms which are stable in proportion to the depth of the minima of the potential curve.

Figure 4.

When the stabilizing forces are weak, a statistical equilibrium is set up between the isomers due to thermal agitation. These isomers are indistinguishable chemically because the potential barriers opposing free rotation are small compared

* Stitt (1941), studying the infra-red spectrum from 400 to 4000 cm.$^{-1}$, has concluded that the structure is analogous to that of ethane.

with the energy liberated in the substitution reactions. However, the relative proportions can be calculated from spectroscopic analysis (Kohlrausch, 1932). The number of lines in the Raman spectrum, and the variation of their intensity with temperature, sometimes suffices to fix completely the number of isomers, and also to show which is the most stable (Kohlrausch and Köppl, 1935; Glockler and Sage, 1941). This is the problem of rotational isomerism.

(i) *Ethane*

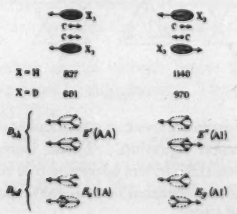
If the two halves of the molecule rotate about the C—C axis with respect to each other, it is not possible to determine directly the torsional frequency

Figure 5.

by analysis of the absorption or Raman spectrum; the torsional vibrations of a molecule like ethane (D_{3h} or D_{3e}) are inactive in absorption and scattering. But the number and classification of the lines absorbed and emitted by the molecules C_2H_6 and C_2D_6 show that each exists in one form only. Thermodynamic studies indicate that on all sides of this stable orientation there is a potential barrier of 2750 cal./mol., from which the torsional frequencies, 275 cm.$^{-1}$ for ethane and 200 cm.$^{-1}$ for deutero-ethane, are deduced. It is seen, then, that rotation is not free, and it is necessary to choose between the models $\alpha(D_{3h})$ and $\beta(D_{3d})$. This choice is not easy.

It is first necessary to classify the observed frequencies in the Raman and infra-red spectra for the molecules C_2H_6 and C_2D_6 (Crawford, Avery and Linnett, 1938; Glockler and Renfrew, 1938; Goubeau and Karweil, 1938; Karweil and Schäfer, 1938; Stitt, 1939). For this, use has been made of selection and polarization rules, of *a priori* calculation of the frequencies from simple potential functions, of the product rule, and of specific-heat measurements. Different authors appear to agree upon the following essential fact: the frequencies 1140

Table 6



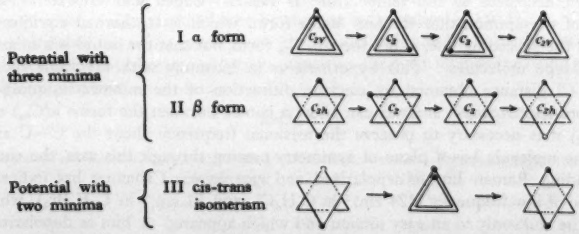
and 827 for ethane and 970 and 601 for deutero-ethane are to be attributed to the degenerate rotations shown in table 6, and the groups D_{3h} and D_{3d} are only distinguishable by the different selection rules for the type of oscillation E' of the group D_{3h}, and the type E_u of the group D_{3d}.

The line 601 is very intense in absorption, but does not appear in the Raman spectrum of deutero-ethane, while for ethane only Glockler and Renfrew have been able to observe a very faint Raman line for the liquid at about 813 cm.$^{-1}$, which can arise from molecules deformed by intermolecular forces. Hence the $\beta(D_{3d})$ model is the form that comes to mind, but it should be noted that the frequency 1140 cm.$^{-1}$ of C_2H_6, which give rise to a Raman line for both models, is not observed. On the other hand the Raman line at 970 for C_2D_6 is not very intense but is distinctly visible. Hence the following conclusion can be drawn: For C_2D_6 the selection rules indicate a D_{3d} symmetry; for C_2H_6 there is no clear indication, but it is obvious that the two isotopic molecules have the same symmetry.

The "uncertain frequency" $\omega = 1140$ for ethane, which does not appear in the optical spectrum, is deduced from the specific heat and the product rule. If this frequency is accepted, then the infra-red frequency 2230 is the harmonic 2ω, just as the frequency 1930 is the harmonic of the frequency 970 for deutero-ethane. Now the selectrion rules for frequencies of the second order (Tisza, 1933) prohibit the frequency 2ω in the group D_{3d}, but allow it for group D_{3h}. For this reason Karweil and his collaborators have adopted the α form, while Stitt, who prefers the β form, has reached no final conclusion.

Langseth and Bernstein (1940) approached the problem indirectly. During an important study on rotational isomerism they noticed, for the first time, that the potential energy of a molecule $CH_2X—CH_2Y$ could have minimum values for configurations other than the cis- and trans- forms then known. They distinguished the three following cases:

Table 7



In the Raman spectrum of the bromine derivative $CH_2D—CH_2Br$, a doubling of the C—Br frequency is observed, the two components being 536·5 and 557·5. The intensity of the line 557 is twice as great as that of the line 536, and the latter increases in intensity while the line 557 decreases with rise in temperature. Hence there are several co-existing molecular forms, and the less stable

molecules are twice as numerous as the stable ones. From this, a potential energy curve having three minima is obtained, which means that the configuration is either α, having the cis- form (C_{2v}) and two C_2 forms, or β, having the trans- form (C_{2h}) and two C_2 forms. By analogy with other isomers it is assumed that the lowest of the C—Br frequencies is due to the cis- form of the molecule; hence the molecule $CH_2D—CH_2Br$ will have form α. The same conclusion is indicated for the molecule $CH_3—CH_2Br$ and the authors adopted it too for ethane.

(ii) *Halogenated ethanes*

(*a*) $CH_2X—CH_2Y$ (X, Y = Cl, Br)

The first example of rotational isomerism was provided by di-halogenated ethanes. The trans- form (C_{2h}) of the molecule $CH_2X—CH_2X$ possesses a centre of symmetry, therefore its Raman and infra-red spectra have no common frequencies, and only the completely symmetrical rotations of the chain $X\!\!\diagup\!C—C\!\!\diagup\!X$ scatter light actively. In the solid state all the molecules have the trans- form, but a study of the liquid reveals that (i) the characteristic frequencies of the chain are too numerous to be all due to the single trans- form; (ii) several Raman lines are depolarized (Cheng, 1935); (iii) certain vibrations are very active both in absorption and scattering (Cheng and Lecomte, 1935). From these Kohlrausch and Ypsilanti (1935) concluded that two molecular forms co-exist in the liquid at ordinary temperatures and that the trans- form is the more stable and the more abundant. Cabannes (1938) and Wu (1939) revised the classification of the frequencies of the molecules $C_2H_4Cl_2$, C_2H_4ClBr and $C_2H_4Br_2$, and showed that all the details of their spectra could readily be explained on the basis of the coexistence of the cis- and trans- forms.

The most stable di-halogenated ethane is undoubtedly the trans- C_{2h} form, which is the only one that occurs in the crystal, and the proportion of it in the liquid decreases as the temperature is raised. Edgell and Glockler (1941) are of the opinion that the less stable form, which is in thermal equilibrium with the preceding one, is not the cis- C_{2v} form, but that the liquid is a mixture of β-type molecules. This hypothesis is in harmony with the value of the Cl—Cl distance obtained by electron diffraction of the mixture in question. Before spectroscopic analysis can allow a choice between the forms $\alpha(C_{2v})$ and $\beta(C_2)$ it is necessary to observe the torsional frequency about the C—C axis. If the molecule has a plane of symmetry passing through this axis, the corresponding Raman line is depolarized, and vice versa. Cabannes has indicated a line of low frequency (124 cm^{-1} in $C_2H_4Cl_2$ and 91 cm^{-1} in $C_2H_4Br_2$), which can be due only to an easy torsion and which appeared to him as depolarized. Although the depolarization was doubtful, Cabannes was able to deduce that the molecule was in the cis- form. Later, Kohlrausch and Wittek (1940) confirmed these two depolarized lines at 124 and 89. But measurement of depolarization is inaccurate for Raman lines in the neighbourhood of a strong line scattered without change of wave-length. Hence we can agree with Edgell and

Glockler that di-halogenated ethanes have three stable β forms, the trans- C_{2h} form being the most stable:

CH_2X-CH_2Y
[X, Y = F, Cl, Br, I]

(*b*) CHX_2—CHX_2 and CDX_2—CDX_2 (X = Cl, Br)

In tetra-halogenated ethanes Langseth and Berstein (1940) found many frequencies under 400 cm.$^{-1}$; each of them could thus not be looked upon as a mixture of β-molecules or as a mixture of cis- and trans- molecules. Therefore the liquid consists of α-type molecules:

CHX_2—CHX_2
CDX_2—CDX_2

(*c*) CX_3—CX_3 (X = Cl, Br)

In this case, as in ethane, it is a question of choosing between the α form (D_{3h}) and the β form (D_{3d}). Analysis of the Raman spectrum led Hamilton and Cleveland (1944) to accept the D_{3d} symmetry.

(iii) *Other examples*

In the last few years these researches have been extended to numerous aliphatic chains, e.g. $OHCH_2$—CH_2OH, CH_3CH_2—CH_2CH_3, CH_3CH_2—CH_2X, XCH_2—CH_2OH, where X = Cl, Br, I. In general only one form is observed in the solid state, while several molecular forms co-exist in the dissolved or liquid state. To give some idea of the methods used, and the conclusions that can be drawn, the work of Kohlrausch and his co-workers on the primary aliphatic derivatives RCH_2CH_2X (X = SH, Cl, Br, I) will be cited. In each case two Raman frequencies ω' and ω'' are observed which are related to the vibrations about the C—X bond. These two frequencies are unaffected by changes in the radical $R = C_nH_{2n+1}$. The line ω' of lower frequency is the more intense and only the iso-butyl derivatives (branched at the β-carbon atom) behave otherwise, the more intense line being that of high frequency ω''.

Table 8

	RCH_2—CH_2X		$(CH_3)_2CH$—CH_2X	
	ω'	ω''	ω'	ω''
X = Cl	652 (7)	723 (4)	684 (6)	726 (10)
= Br	562 (6)	643 (4)	631 (6)	651 (10)
= I	504 (8)	594 (6)	582 (6)	602 (9)

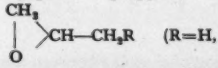
The frequencies ω' and ω'' of the derivatives $RCH_2 . CH_2X$ are very close to those of the di-halogenated derivatives $XCH_2 . CH_2X$, which are 652 and 752

for X = Cl and 553 and 659 for X = Br; so it is assumed that the molecular forms present are of the type

If it is assumed that the iso-butyl derivatives have the forms

the reason for the line ω'' being more intense can be understood.

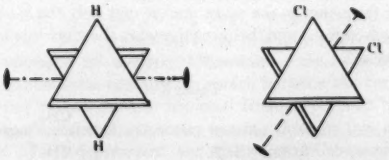
The Raman spectra of propylene oxide and its derivatives,



$(R=H, OH, NH_2, Cl, Br, I, CN, CO_2H)$,

show that these substances have similar structure to iso-butyl derivatives (Ballaus and Wagner, 1939). Resolution of most of the frequencies of the chain (at least in the halogen derivatives) reveals the presence of several co-existing molecular forms. Before a final conclusion can be reached it is necessary to study an additional large number of analogous structures, and make a close critical study of all the experimental data.

If the spectra of the propyl derivatives $CH_3 . CHX . CH_3$ and $CH_3 . CH_2 . CH_2X$ are compared it is found, by classification and deduction from the Raman spectra, that the carbon chain of the secondary derivative is rigid, while that of the primary derivative is mobile. The same difference is noticed for iso-propylene derivatives $CH_2 = CX . CH_3$ (rigid chain) and allyl derivatives $CH_2 = CH—CH_2X$ (mobile). The allyl derivatives exist in two forms in the liquid state (Kahovec and Kohlrausch, 1940).



Butadiene Hexachlorobutadiene

Figure 6.

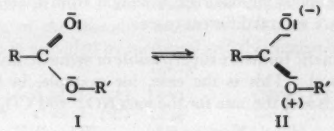
The case of butadiene is still more typical. The scattering spectrum was analysed under high dispersion and the depolarization factor measured (Bradacs and Kahovec, 1940). All the details of the spectrum are adequately explained if a C_{2h} form, rigid enough to remain unaltered by dissolution in a variety of solvents, is assumed. In contrast to butadiene the Raman spectrum of the melted hexachlorobutadiene C_4Cl_6 shows that the carbon chain of this molecule possesses a non-planar symmetry (figure 6).

To sum up, when two parts of the same molecule can change their respective positions by rotating about their bond, spectroscopic analysis shows that at ordinary temperatures the rotation is not continuous, as would be indicated by the spreading of Raman lines, but that a small number of well-defined molecular forms co-exist in the liquid, each characterized by its own particular spectrum.

§ 7. MESOMERISM

From the study of chemical reactions it is found that for unsaturated substances, different formulae can be written and that all of them are equally, or almost equally, plausible. When the arrangment of the nuclei remains the same, and the formulae differ only in electronic configuration, the phenomenon is called mesomerism. If the octet principle is respected, formulae are obtained of which two examples are given below:

Esters



I II

Ionized salt

I II

In these formulae each dash represents two electrons. According as the dash is allotted to one atom alone or joins it to its neighbour, so the two electrons belong to that atom alone or is shared by them. The octet principle demands four lines to each nucleus. Without changing anything in the hypothesis, these formulae can be written more simply thus:

Esters

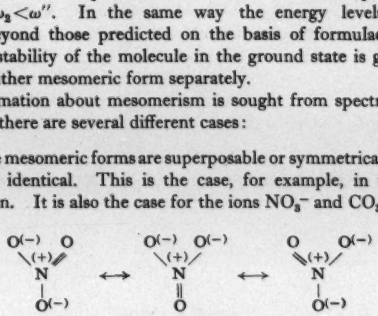
Ionized salt

In these formulae only the lines which represent bonds have been retained, and the signs (+) or (−) above an atom indicate that it carries a positive or negative charge.

For the ionized salt forms I and II are equally likely, but for esters the molecule will be less likely to possess form II than form I, because the introduction of a dipole increases its potential energy. In actual fact the observed properties, e.g. heat of formation, interatomic distance, etc., deviate more or less from those which would be expected from one or other of these formulae, and wave mechanics indicates that neither truly represents the molecule. It can be imagined that all the mesomeric forms together bear the same relation to the true molecule as a system of uncoupled electric oscillators bears to the same system after the coupling. To the two frequencies ω_1 and ω_2 of the two isolated oscillators there correspond two frequencies ω' and ω'' of the coupled system such that $\omega' < \omega_1 < \omega_2 < \omega''$. In the same way the energy levels of the actual molecule lie beyond those predicted on the basis of formulae I and II. In particular, the stability of the molecule in the ground state is greater than that calculated for either mesomeric form separately.

When information about mesomerism is sought from spectroscopic analysis it is found that there are several different cases:

(i) The mesomeric forms are superposable or symmetrical, and, therefore, chemically identical. This is the case, for example, in the ions RCO_2^- already seen. It is also the case for the ions NO_3^- and CO_3^{--}.



The real electronic configuration depends upon each of the mesomeric forms to the same extent, and the mechanism which allows one form to pass into another appears to be a covering operation for the electronic field and arrangement of the nuclei. In this way a binary axis has been found in the $-CO_2$ radical of the acetate ion, and in the $-NO_2$ radical of nitroparaffins, a three-fold axis in NO_3 and CO_3 ions, and a six-fold axis in the benzene molecule. Thus the symmetry of the molecule is higher than that of the mesomeric forms which enter into resonance, and by showing the symmetry of the arrangement of the nuclei, spectroscopic analysis is sufficient to show that mesomerism is taking place.

(ii) If one of the possible molecular forms already possesses the observed symmetry, the rules of selection and polarization do not show whether mesomerism takes place, nor to what extent. For example, these three linear forms can be envisaged for gaseous carbon disulphide and carbon dioxide

I	II	III
$O{=}C{=}O$	$O^{(-)}{-}C^{(+)}{\equiv}O$	$O^{(+)}{\equiv}C^{(-)}{-}O$
$S{=}C{=}S$	$S^{(-)}{-}C^{(+)}{\equiv}S$	$S^{(+)}{\equiv}C^{(-)}{-}S$

Whether form I exists alone or enters into resonance with forms II and III, the symmetry $D_{\infty h}$ will be observed, but measurement of the frequencies will provide the information required (Duchesne, 1943). If, in a linear

molecule XYZ, the displacements of the nuclei along the axis which joins them are designated by ξ_1, ξ_2, ξ_3, the potential energy is expressed as

$$V = \tfrac{1}{2}[F(\xi_2-\xi_1)^2 + F'(\xi_3-\xi_2)^2 + f(\xi_2-\xi_1)(\xi_3-\xi_2)] = \tfrac{1}{2}[F\rho^2 + F'\rho'^2 + f\rho\rho'].$$

When the molecule has the symmetry $D_{\infty h}(F' = F,\ m_3 = m_1)$, the formulae

$$2F = m_1\omega_1^2 + \mu\omega_2^2, \qquad 2f = m_1\omega_1^2 - \mu\omega_2^2,$$

where

$$\frac{1}{\mu} = \frac{1}{m_1} + \frac{2}{m_2},$$

give the resistance to deformation F and the interaction of the bonds f. The frequencies ω_1 and ω_2 are the fundamental vibration frequencies along the axis. For the molecules CO_2 and CS_2 they are

CO_2	$\omega_1 = 1337\ \text{cm.}^{-1}$	$\omega_2 = 2349$	$10^{-5}F = 15{\cdot}45$ dynes/cm.
CS_2	$\omega_1 = 655{\cdot}5$	$\omega_2 = 1523$	$10^{-5}F = 7{\cdot}47$

Now the value of $10^{-5}F$ for double and triple bonds is known from the band spectra of the diatomic molecules CO and CS, and the value for single bonds is provided by methanol and thiomethanol:

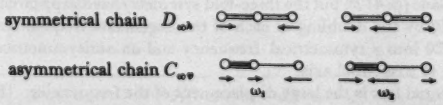
	$10^{-5}F$		$10^{-5}F$
$C \equiv O$ $(^1\Sigma)$	18·92	$C \equiv S$ $(^1\Sigma)$	8·46
$C = O$ $(^3\Pi)$	12·16	$C = S$ $(^3\Pi)$	5·87
$H_3C - OH$	5·3	$H_3C - SH$	3·6
Average	12·19	Average	5·98

In the molecules CO_2 and CS_2 the bonds C—O and C—S are intermediate between double and triple bonds which have been interpreted by Pauling and Sherman (1933) as being due to resonance, the actual molecule being more closely knit than any of the forms proposed above.

(iii) In the third case, by far the most common, the mesomeric forms, are all chemically distinct. Then one is compelled to admit *a priori* mesomerism as a working hypothesis, and to investigate whether it suffices to explain the anomalies of different kinds that occur in the position and intensity of spectral lines. It is in this way that Kohlrausch and his collaborators worked for the last few years, and the validity of the results rests upon the more or less large number of observed facts (Kohlrausch, 1942). Some recent examples will demonstrate the method and the results which can be obtained.

(i) *Examples of mesomerism in linear triatomic chains*

Let ω_1 and ω_2 be the frequencies of vibration of the nuclei along the chain. If the masses of the nuclei are not very different the vibrations can be represented thus:



When the chain is symmetrical, only the frequency ω_1 is observed in the Raman spectrum. When it is asymmetrical both frequencies appear in it and the

vibration ω_1, in the course of which the tensor [α] undergones larger variations, gives a more intense line than does the vibration ω_2.

So much having been stated, we will now compare the results obtained from thirteen molecules. Table 9 shows the most reasonable mesomeric forms of each molecule and the frequencies of the Raman lines, with an indication of their intensity (ff = very weak, f = weak, m = medium, F = strong).

The table shows that the ion N_3^- has $D_{\infty h}$ symmetry, while the acid N_3H has $C_{\infty v}$ symmetry. On the other hand, the frequencies of the acid are not affected by substituting a CH_3 radical for the H atom. The fundamental change of the chain produced by passing from the N_3^- to the N_3H acid, or the ester $N_3(CH_3)$, is not found on passing from allene (1) to methyl allene (2). The frequency ω_2 is unobservable for methyl allene in which the $D_{\infty h}$ symmetry of the chain C = C = C is preserved. These facts are explained by mesomerism. It does not take place in allene, and when allene is methylated the electronic configuration of the molecule remains unchanged. On the contrary, the combination of the mesomeric forms of N_3^- has $D_{\infty h}$ symmetry, while the combination of the mesomeric forms of the acid has $C_{\infty v}$ symmetry, and this is confirmed by experience.

The ion NCO^- (7) in potassium cyanate has $C_{\infty v}$ symmetry; it is, therefore, form I which predominates. This appears to result from the fact that the oxygen nucleus has a larger positive charge than the nitrogen nucleus and consequently has a greater attraction for electrons. The Raman spectrum of the acid only shows the line ω_1. Therefore this acid is the isocyanic (9), of which the most stable form, HN = C = O, is pseudo-symmetrical, while the form having the least energy for cyanic acid, N≡C—OH (6), is asymmetrical. The increased value (3297) of the frequency of vibration of the hydrogen nucleus in the direction of the bond confirms that it is bound to the nitrogen atom. The pseudo-symmetry of the acid is preserved in methyl iso-cyanate (10).

When a sulphur atom replaces the oxygen atom in O = C = N, the symmetry of the chain vanishes because of the mass effect (13). Both lines, ω_1 and ω_2, are observed in the Raman spectrum of (11), (12) and (13). The frequencies and their corresponding intensities confirm the predominance of form I in thiocyanic (11), and of form II in iso-thiocyanic esters (13). Form I also predominates in the ions (12) and (14), which seems to indicate, as in example (7), that the electron is attracted to the nucleus with the greatest positive charge.

(ii) *Triangular aggregates*: NO_2 *and* CO_2

In the ions CO_3^{--} and NO_3^- the D_{3h} symmetry of the electronic configuration is shown by the symmetrical arrangment of the nuclei, which gives rise to equally distributed forces between the central nucleus and those of oxygen. When the ion NO_3^- passes into the acid NO_3H, the NO_3 grouping remains planar with a binary axis in the plane (p. 412), but the three-fold symmetry has disappeared, as is shown (see table 10) by the doubling of each of the degenerate frequencies $\omega_{3,4} = 1360$ and $\omega_{3,5} = 720$ into a symmetrical frequency and an antisymmetrical frequency with respect to a two-fold axis.

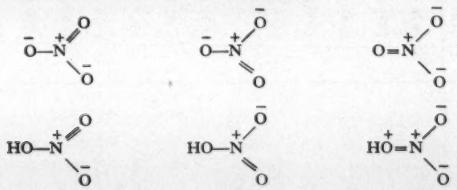
The essential fact is the large displacement of the frequencies. For example, the frequency ω_1 passes from 1050 to 1296, showing that strength of one of the N—O bonds is decreased and is not more than half of each of the other two.

Table 9

	I	II	III	Raman lines	
				ω_1	ω_2
1	$CH^2{=}CH{=}CH^2$	—	—	1074 F	—
2	$(CH_3)CH{=}C{=}CH_2$	—	—	1083 F	—
3	$N{\equiv}\overset{+}{N}{-}\overset{-}{O}$	$\overset{-}{N}{=}\overset{+}{N}{=}O$	—	1286 F	2223 f
4	$N{\equiv}\overset{+}{N}{-}\overset{-}{N}H$	$\overset{-}{N}{=}\overset{+}{N}{=}NH$	—	1300 m	2389 f
5	$N{\equiv}\overset{+}{N}{-}\overset{--}{N}$	$\overset{-}{N}{=}\overset{+}{N}{=}\overset{-}{N}$	$\overset{--}{N}{-}\overset{+}{N}{\equiv}N$	1348 F	—
6	$[N{\equiv}C{-}OH]$	$[\overset{-}{N}{=}C{=}\overset{+}{O}H]$	—	—	—
7	$N{\equiv}C{-}\overset{-}{O}$	$\overset{-}{N}{=}C{=}O$	—	857 f	2192 F
8	← Mercury and silver iso-cyanates. →			1302 F	2206 f
9	$H\overset{+}{N}{\equiv}C{-}\overset{-}{O}$	$HN{=}C{=}O$	$H\overset{-}{N}{-}C{\equiv}\overset{+}{O}$	1307 F	—
10	$(CH_3)\overset{+}{N}{\equiv}C{-}\overset{-}{O}$	$(CH_3)N{=}C{=}O$	$(CH_3)\overset{-}{N}{-}C{\equiv}\overset{+}{O}$	1409 F	—
11	$N{\equiv}C{-}S(CH_3)$	$\overset{-}{N}{=}C{=}\overset{+}{S}(CH_3)$	—	697 m	2153 F
12	$N{\equiv}C{-}\overset{-}{S}$	$\overset{-}{N}{=}C{=}S$	—	750 f	2066 F
13	$(CH_3)\overset{+}{N}{\equiv}C{-}\overset{-}{S}$	$(CH_3)N{=}C{=}S$	$(CH_3)\overset{-}{N}{-}C{\equiv}\overset{+}{S}$	1087 m	2162 f
14	$N{\equiv}C{-}\overset{-}{Se}$	$\overset{-}{N}{=}C{=}Se$	—	575 f	2051 m

Hence, in the NO_2H molecule a NO_2 grouping can be identified which has three internal vibrations and two "external" ones. The two parts NO_2 and OH of the molecule are relatively displaced during the latter. The frequency $\omega_{2,4}$ is doubled, giving an antisymmetrical internal frequency $\omega_4 = 1683$ and an external symmetrical frequency $\omega_2 = 920$; the frequency $\omega_{3,5}$ is separated into a symmetrical internal frequency $\omega_3 = 608$ and an external antisymmetrical one $\omega_5 = 674$.

This modification of the intermolecular forces can be caused only by important changes in the electronic configuration:

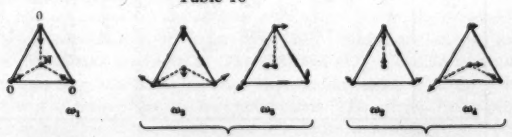


I II III

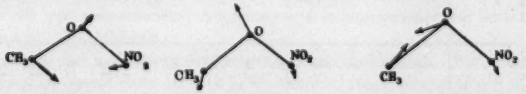
In the ion the mesomeric forms I, II and III play the same part, but in the acid form III plays a smaller part, and so the bond N—OH is only slightly different from a single bond.

On passing from the acid to esters by substitution of a chain C_nH_{2n+1} for the hydrogen atom of the acid, only slight modifications of the internal frequencies are observed, showing that the electronic configuration of the grouping NO_2 does not change appreciably, but frequencies corresponding to an angular grouping of NOC appear in methyl nitrate.

Table 10

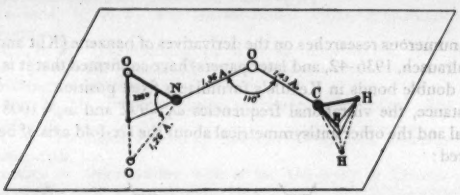


NO_3^-	1050 P		720 D		1360 D	
CO_3^{--}	1070 P		710 D		1430 D	
NO_2—OH	1296 P		608 P	674 D	920 P	1683 D
NO_2—O—CH_3	1291 P	355 P	582 P	668 D	865 P 996 P	1644 D
CH_3—NO_2	1375 P	483 D	610 P	657 D	918 P	1562 D
CH_3—$CO_3]^-$	1347 P	475 D	613 P	650 D	927 P	1542 D



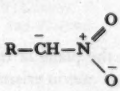
ω =	355	865	996
ρ =	0·45	0·26	0·68

Lecomte and Mathieu (1942) have assigned to methyl nitrate the structure shown below, in which the plane containing NO_2 is perpendicular to the plane containing the N—O—C chain, because the Raman lines 668 and 1644 are depolarized as in the acid.

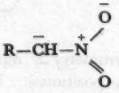


Further, table 10 shows the resemblance between nitromethane and the acetate ion. The NO_2 radical of the nitro-paraffins differ less from the CO_2^- radical than from the NO_2 radical of nitro-esters. The molecule $CH_3—NO_2$ and the ion $(CH_3—CO_2)^-$ have the same electronic configuration, so have the monovalent ion NO_2 and the divalent ion CO_3^{--}. The additional electrons which give the anions their charge set up an external electric field without modifying the internal structure.

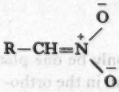
Nitro-paraffins give alkali salts, and one can postulate, for the anion, three mesomeric forms :



I



II

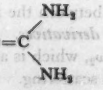


III

But a strong line near 1580 is observed in the Raman spectrum. This line is polarized, and is, therefore, not of the same origin as the depolarized lines of the nitro-paraffins. It arises from a double C = N bond, which confirms the formula III attributed to the anions of these salts. Likewise the line 1660 for urea is polarized. This arises from a double C = O bond. Of the three mesomeric forms

I

II



III

it is, therefore, still the third which has the predominating influence.

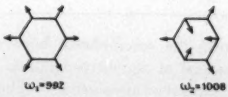
(iii) *The benzene nucleus*

Since the work of Angus, Bailey, Hale, Ingold, Leckie, Raisin, Thomson and Wilson (1935), it has not been doubted that the benzene molecule

possesses D_{6h} symmetry, and it is assumed that this high symmetry is essentially the result of resonance between the two Kekule forms:

Since then numerous researches on the derivatives of benzene (Klit and Langseth, 1937; Kohlrausch, 1936–42, and later papers) have confirmed that it is impossible to give the double bonds in Kekule's formulae a fixed position.

For instance, the vibrational frequencies $\omega_1 = 992$ and $\omega_2 = 1008\,\text{cm.}^{-1}$, one symmetrical and the other antisymmetrical about the six-fold axis of benzene, will be considered:



The frequency ω_2 does not appear in the Raman spectrum. Now suppose that two hydrogen atoms be replaced by two deuterium atoms or any two radicals X. If the double bonds are supposed to be fixed,

there can only be one plane of symmetry Σ' normal to the plane of the hexagon (substitution in the ortho- or para- positions). If there is mesomerism there may be two orthogonal planes of symmetry Σ and Σ' (substitution in the para-position).

One is therefore led to the following conclusions:

Ortho- derivatives. In both hypotheses the molecule has a symmetry plane Σ', and the vibrations ω_1 and ω_2 are respectively symmetrical and antisymmetrical with relation to this plane.

Meta- derivatives. The Σ' planes does not exist in either hypothesis, and the vibrations ω_1 and ω_2 are of the same type. Besides, as their frequencies are close to each other, the anharmonicity gives rise to a coupling which increases the difference between the frequencies.

Para- derivatives. In the second hypothesis there is a centre and the vibration ω_2, which is antisymmetrical with reference to this centre, ought to be inactive in scattering. This is confirmed by experiment. This vibration would have given a Raman line in the first hypothesis.

Hence the study of the para- disubstituted compounds indirectly confirms the six-fold symmetry of the benzene nucleus.

The author would like to thank M. Robert Lennuier, agrégé des sciences physiques, assistant in the Physical Research Laboratories in the Sorbonne, who has given much help in the preparation of this paper.

GENERAL REFERENCES

1931 : WIGNER, E. *Gruppentheorie* (Braunschweig).

1934 : PLACZEK, G. *Rayleigh Streuung und Raman-Effekt. Handbuch den Radiologie*, IV/2 (Leipzig : Akademische Verlagsgesellschaft).

1936 : ROSENTHAL, J. E. and MURPHY, G. M. Group Theory and the Vibrations of Polyatomic Molecules. *Rev. Mod. Phys.*, **8**, 317.

1938 : KOHLRAUSCH, K. W. F. *Der Smekal-Raman Effekt, Ergänzungsband* (Berlin : J. Springer).

1938 : RANDALL, H. M. The Spectroscopy on the far Infra-red. *Rev. Mod. Phys.*, **10**, 72.

1939 : HIBBEN, J. H. *The Raman Effect and its Application* (New York : Reinhold Publishing Corporation).

1940 : DENNISON, D. M. Infra-red Spectra of Polyatomic Molecules. *Rev. Mod. Phys.*, **12**, 175.

1941 : WU, T. Y. *Vibrational Spectra and Structure of Polyatomic Molecules* (New York : Prentice-Hall, Inc.).

1942 : CONFERENCE ON SPECTROSCOPY, held at the University of Chicago. *Rev. Mod. Phys.*, **14**, 55.

1943 : CABANNES, J. Les spectrographes modernes et leurs applications à la physique moléculaire. *Revue Scientifique*, **81**, 419.

1943 : GLOCKLER, G. The Raman Effect. *Rev. Mod. Phys.*, **15**, 11.

1943 : KOHLRAUSH, K. W. F., SIMON, A. and GOUBEAU, J. Raman Effekt und Konstitution. Anorganische Verlindungen. Organische Moleküle. *Z. Electrochemie*, **49**, 407 and 431.

1943 : KOHLRAUSCH, K. W. F. *Ramanspektren* (Leipzig : Becker und Erler).

OTHER REFERENCES

ANANTHAKRISHNAN, R., 1937. *Proc. Ind. Acad. Sci., Bangalore*, **5**, 76, 87 and 200.

ANGUS, W. R., BAILEY, C. R., HALE, J. B., INGOLD, C. K., LECKIE, A. H., RAISIN, C. G., THOMSON, J. W. and WILSON, C. L., 1936. *J. Chem. Soc.*, 971.

BAILEY, C. R., CARSON, S. C. and DALY, E. F., 1939. *Proc. Roy. Soc.*, **A, 173**, 339.

BALLAUS, O. and WAGNER, J., 1939. *Z. phys. Chem.*, B **45**, 165.

BARKER, E. F., 1942. *Rev. Mod. Phys.*, **14**, 198.

BARNES, R. B., 1934. *Rev. Sci. Instrum.*, **5**, 237.

BARNES, R. B., 1935. *Phys. Rev.*, **47**, 658.

BRADACS, K. and KAHOVEC, L., 1940. *Z. phys. Chem.*, B **48**, 63.

BRESTER, C. J., 1923. *Dissertation*, Utrecht.

CABANNES, J., 1932. *Ann. Phys., Paris*, **18**, 285.

CABANNES, J., 1938. *J. Chim. phys.*, **35**, 1.

CABANNES, J., 1942. *Rev. Sci., Paris*, **80**, 60 and 407.

CABANNES, J. and DUFAY, J., 1935. *C.R. Acad. Sci., Paris*, **200**, 1504.

CABANNES, J. and ROUSSET, A., 1940. *J. Phys. Radium*, **1**, 155, 181 and 210.

CAMERON, D. M., SEARS, W. C. and NIELSEN, H. H., 1939. *J. Chem. Phys.*, **7**, 994.

CHÉDIN, J., 1937. *Ann. Chim.* (*Phys.*), **8**, 243.

CHÉDIN, J., 1939. *J. Phys. Radium*, **10**, 445.

CHENG, H. C., 1935. *J. Chim. phys.*, **32**, 715.

CHENG, H. C. and LECOMTE, J., 1935. *J. Phys. Radium*, **6**, 477.

CLEVELAND, F. F., 1943. *J. Chem. Phys.*, **11**, 1.

COBLENTZ, W. W., 1920. *J. Opt. Soc. Amer.*, **4**, 432.

COJAN, J., 1932 and 1933. *Réunions de l'Institut d'Optique* (Paris : Editions de la Revue d'Optique).

CONRAD-BILLROTH, H. C., KOHLRAUSCH, K. W. F. and REITZ, A. W., 1937. *Z. Electrochem.*, **43**, 293.

CRAWFORD, B. L., JR., AVERY, W. H. and LINNET, J. W., 1938. *J. Chem. Phys.*, **6**, 682.

CZERNY, M. and MOLLET, P., 1937. *Phys. Z.*, **38**, 1008.

CZERNY, M. and MOLLET, P., 1937. *Z. tech. Phys.*, **18**, 582.

CZERNY, M. and MOLLET, P., 1938. *Z. Phys.*, **108**, 85.

DELWAULLE, M. L., 1943 a. *C.R. Acad. Sci., Paris*, **216**, 735.

DELWAULLE, M. L. 1943 b. *C.R. Acad. Sci., Paris*, **217**, 172.

DELWAULLE, M. L. and FRANÇOIS, F., 1941. *C.R. Acad. Sci., Paris*, **212**, 761.
DELWAULLE, M. L. and FRANÇOIS, F., 1942. *C.R. Acad. Sci., Paris*, **214**, 226 and 828.
DELWAULLE, M. L. and FRANÇOIS, F., 1944. *C.R. Acad. Sci., Paris*, **219**, 64 and 335.
DELWAULLE, M. L. and FRANÇOIS, F., 1945. *C.R. Acad. Sci., Paris*, **220**,
DENNISON, D. M., and HARDY, J. D., 1932 *Phys. Rev.*, **39**, 938.
DENNISON, D. M., and UHLENBECK, G. E., 1932. *Phys. Rev.*, **41**, 313.
DIEKE, G. H. and CROSSWHITE, H. M., 1945. *J. Opt. Soc. Amer.*, **35**, 471.
DUCHESNE, J., 1943. *Thesis*, Liège.
EDGELL, W. F. and GLOCKLER, G., 1941. *J. Chem. Phys.*, **9**, 375 and 484.
FIRESTONE, F. A., 1932. *Rev. Sci. Instrum.*, **3**, 163.
FOLEY, H. M. and RANDALL, H. M., 1941. *Phys. Rev.*, **59**, 171.
GALLAWAY, W. S. and BARKER, E. F., 1942. *J. Chem. Phys.*, **10**, 88.
GLOCKLER, G. and EDGALL, W. F., 1941. *J. Chem. Phys.*, **9**, 224 and 527.
GLOCKLER, G. and LEADER, G. R., 1940. *J. Chem. Phys.*, **8**, 125 and 699.
GLOCKLER, G. and RENFREW, M. M., 1938. *J. Chem. Phys.*, **6**, 408 and 409.
GLOCKLER, G. and SAGE, C. G., 1941. *J. Chem. Phys.*, **9**, 387.
GOUBEAU, J., 1935. *Ber. dtsch. chem. Ges.*, **68**, 912.
GOUBEAU, J. and KARWEIL, J., 1938. *Z. phys. Chem.*, B **40**, 376.
HAMILTON, D. T. and CLEVELAND, F. F., 1944. *J. Chem. Phys.*, **12**, 249.
HARDY, J. D., 1930. *Rev. Sci. Instrum.*, **1**, 429.
HOPF, M., 1940. *Z. Phys.*, **116**, 310.
KAHOVEC, L. and KOHLRAUSCH, K. W. F., 1940. *Z. phys. Chem.*, B **46**, 165.
KAHOVEC, L. and WAGNER, J., 1939. *Z. phys. Chem.*, B **42**, 123.
KAHOVEC, L. and WAGNER, J., 1940. *Z. phys. Chem.*, B **47**, 48.
KARWEIL, J. and SCHÄFER, K., 1938. *Z. phys. Chem.*, B **40**, 382.
KASTLER, A. and ROUSSET, A., 1941. *J. Phys. Radium*, **2**, 49.
KLIT, A. and LANGSETH, A., 1937. *J. Chem. Phys.*, **5**, 925.
KOHLRAUSCH, K. W. F., 1932 a. *Z. phys. Chem.*, B **18**, 61.
KOHLRAUSCH, K. W. F., 1932 b. *Z. phys. Chem.*, B **20**, 217.
KOHLRAUSCH, K. W. F., 1936. *Phys. Z.*, **37**, 58.
KOHLRAUSCH, K. W. F., 1942. *Öst. ChemZtg.*, **45**, 169.
KOHLRAUSCH, K. W. F. and KÖPPL, F., 1935. *Wien. Ber.*, **143**, 537.
KOHLRAUSCH, K. W. F. and SEKA, R., 1938. *Ber. dtsch. chem. Ges.*, **71**, 985, 1551 and 1563.
KOHLRAUSCH, K. W. F. and SEKA, R., 1940. *Ber. dtsch. chem. Ges.*, **73**, 162.
KOHLRAUSCH, K. W. F. and STOCKMAIR, W., 1935. *Z. phys. Chem.*, B **29**, 292.
KOHLRAUSCH, K. W. F. and WAGNER, J., 1939. *Z. phys. Chem.*, B **45**, 93.
KOHLRAUSCH, K. W. F. and WAGNER, J., 1942. *Z. phys. Chem.*, B **52**, 185.
KOHLRAUSCH, K. W. F. and WITTEK, H., 1940. *Z. phys. Chem.*, B **47**, 55.
KOHLRAUSCH, K. W. F. and WITTEK, H., 1941. *Z. phys. Chem.*, B **49**, 163.
KOHLRAUSCH, K. W. F., and YPSILANTI, G. P., 1935. *Z. phys. Chem.*, B **29**, 274.
KOHLRAUSCH, K. W. F. and YPSILANTI, G. P., 1936. *Z. phys. Chem.*, B **32**, 407.
KUMPF, K. and MECKE, R., 1939. *Z. phys. Chem.*, B **44**, 299.
LANGSETH, A. and BERSTEIN, H. J., 1940. *J. Chem. Phys.*, **8**, 410, 415 and 430.
LECOMTE, J., 1929. *C.R. Acad. Sci., Paris*, **189**, 155.
LECOMTE, J., 1936. *C R. Acad. Sci., Paris*, **203**, 164.
LECOMTE, J., 1938. *C.R. Acad. Sci., Paris*, **196**, 1011.
LECOMTE, J. and MATHIEU, J. P., 1942. *J. Chim. phys.*, **39**, 57.
LEE, E. and WU, C. K., 1939. *Trans. Faraday Soc.*, **35**, 1366.
MATHIEU, J. P. and MASSIGNON, D., 1941. *Ann. Phys., Paris*, **16**, 5.
MAUGUIN, CH., 1931. *Z. Phys.*, **76**, 541.
MENZIES, A. C. and MILLS, H. R., 1935. *Proc. Roy. Soc.*, A, **148**, 407.
NIELSEN, J. R. and WARD, N. E., 1942. *J. Chem. Phys.*, **10**, 81.
PAULING, L. and SHERMAN, J. 1933. *J. Chem. Phys.*, **1**, 606.
PICKARSKI, G. and RUSKA, H., 1939. *Klin. Wschr.*, **18**, 383.
PLACZEK, G., 1931. *Leipziger Vorträge*, No. 71,
RANDALL, H. M., DENNISON, D. M., GINSBERG, N. and WEBER, L. R., 1937. *Phys. Rev.*, **52**, 160.

RANDALL, H. M. and STRONG, J., 1931. *Rev. Sci. Instrum.*, **2**, 585.
RANK, D. H., PFISTER, R. J. and COLEMAN, P. D., 1942. *J. Opt. Soc. Amer.*, **32**, 390.
RANK, D. H., PFISTER, R. J. and GRIMM, H. H., 1943. *J. Opt. Soc. Amer.*, **33**, 31.
REDLICH, O., 1935. *Z. phys. Chem.*, B **28**, 371.
REDLICH, O. and collaborators, 1936. *Wien. Ber.*, **145**, 67, 77, 268 and 328.
REDLICH, O. and NIELSEN, L. E., 1943. *J. Amer. Chem. Soc.*, **65**, 654.
REITZ, A. W., 1940. *Z. phys. Chem.*, B **46**, 181.
ROUSSET, A., LAVAL, J. and LOCHET, R., 1943. *C.R. Acad. Sci., Paris*, **216**, 886.
ROUSSET, A. and LOCHET, R., 1942. *J. Phys. Radium*, **3**, 146.
ROUSSET, A. and LOCHET, R., 1945. *J. Phys. Radium*, **6**, 57.
STITT, F., 1939. *J. Chem. Phys.*, **7**, 297.
STITT, F., 1941. *J. Chem. Phys.*, **9**, 780.
STRONG, J., 1932. *Rev. Sci. Instrum.*, **3**, 810.
SUTHERLAND, G. B. B. M., LEE, E. and WU, C. K., 1939. *Trans. Faraday Soc.*, **35**, 1373.
TELLER, E., after ANGUS, W. R. and others, 1936. *J. Chem. Soc.*, 971.
TISZA, L., 1933. *Z. Phys.*, **82**, 48.
TRUMPY, B., 1931. *K. norske vidensk. Selsk. Forh.*, **4**, 1903.
WAGNER, J., 1941. *Z. phys. Chem.*, B **48**, 309.
WAGNER, J., 1943. *Z. phys. Chem.*, B **53**, 85.
WIGNER, E., 1930. *Nachr. Ges. Wiss. Gottingen, Math.-Phys. Kl.*, 133.
WILSON, E. B., Jr., 1939. *J. Chem. Phys.*, **7**, 1047.
WILSON, E. B., Jr., 1941. *J. Chem. Phys.*, **9**, 76.
WILSON, E. B., Jr., 1943. *J. Chem. Phys.*, **11**, 361.
WITTEK, H., 1942. *Z. phys. Chem.*, B **51**, 187.
WITTEK, H., 1942. *Z. phys. Chem.*, B **52**, 153 and 315.
WRIGHT, N. and RANDALL, H. M., 1933. *Phys. Rev.*, **44**, 391.
WU, T. Y., 1939. *J. Chem. Phys.*, **7**, 965.

LESSONS OF THE WAR FOR SCIENCE*

By J. D. BERNAL, F.R.S.

We have heard much in the last few months of the contributions of science to the war. More than in any previous conflict scientific devices and scientific methods have been employed with success, and, in the end, with decisive success. The gift has not been one-sided ; particularly in the democratic countries, the total voluntary mobilisation of science, for a purpose which scientists recognised as a valid one, may be found in the long run to have enriched the content and method of science quite as much as it contributed to the defeat of forces which would have made science for ever impossible.

There is a very cogent reason why the advancement of science should tend to be associated with the prosecution of war. In human history up to now, it is only in war that the whole energy of a country is turned to one definite end—the defeating of the enemy—and it is particularly in war that this end can be achieved most effectively by the analysis of new situations and the introduction of new methods, operations which call for the techniques and equipment of science. Accordingly, in war, the economic check which limits the expenditure of effort, of discovery and achievement is lifted, and the potentialities of human intelligence can express themselves without the usual traditional checks. All these things are bound to have a stimulating effect on the body of science itself.

If it were not for this, it might well be that the interruption of the ordinary studies of peace-time science, the removal of research workers from their customary laboratories, and the cessation of normal international exchanges, would have interrupted the continuity of scientific research to a possibly irreparable extent. Far from this happening, scientific effort flourished and scientists were fully employed, if not in their usual branch of science, in most cases on some other branch which provided them with experience, both outside and in science, which they would never have obtained otherwise.

The contributions of war to science were both direct and indirect. Many of the researches carried out with an ultimate military purpose necessarily involved intense fundamental research, leading to results which are of the same validity as those which might have been carried out more slowly or, perhaps, not at all in peace time. The most obvious examples are, of course, the contributions made to nuclear physics by the work for the atomic bomb ; the contribution to electronics and the development of radar ; the much greater knowledge of disease and epidemic control through the use of penicillin and D.D.T. ; and innumerable advances in the fields of synthetic chemistry, notably, in the production of synthetic rubber and plastics. It is to be hoped that these scientific results of war research, carried out both in Allied and Axis countries, will be

* The substance of a Friday Evening Discourse given at the Royal Institution on 23 November, 1945.

fully collated and published without delay; but, even without that, the experience gained by the people who took part in them is bound to have a fructifying effect on the progress of science.

As an example in my own experience, I can cite the value to science of the work carried out in this country, in France and in the United States on research into ordinary explosives and their effects. The study of shock and blast waves, which was undertaken in order to understand explosive mechanisms, involved much real fundamental mathematical and physical research. The shock wave, with its intense localised pressure, is an example which transcends the ordinary assumptions involved in the small-amplitude waves of sound. Temperature and pressure effects of explosive shock-waves went beyond what is ordinarily studied in the laboratory and, in certain cases, approached the conditions of stellar atmospheres. The study of the reactions that occur in explosives led, for example, to the formulation of a quantum-mechanical theory in which all the properties of the explosive could be deduced from a knowledge of the quantum levels and pressure properties of a number of simple gases. At least equally important was the entirely new knowledge of the transmission of plastic waves in solid media, and of hydrodynamic disturbances in liquids, both of which are likely to be of the greatest use to basic engineering theory. In the process of studying its effects experimentally, a number of new instruments have been developed which will be of the greatest value in the analysis of all high-speed phenomena. From this example can be seen what a wide range of experimental and theoretical information can be derived from the study of a single practical topic. Such examples could be multiplied in nearly every scientific field. When we remember, also, that all these researches were carried out under the guidance of the most expert theoretical and experimental physicists in their field, with resources such as none of them had ever possessed before, it will be seen that the nett contribution of the war to the actual content of science has not been negligible.

If this were all, however, it might still be argued that more could have been achieved by the same people if they had been allowed to continue their peace-time researches in the ordinary way. Less details might have been found out about things that were already known in principle, but there would always have been the chance of entirely new discoveries which, in the concentration on war, either would not be made or could not be followed up. This argument would be very difficult to sustain if we considered also the greater experimental facilities available, and the stirring up of the minds, even of senior scientists, by contact with unusual problems. But, in any case, the direct contribution of war science to peace science is only a very small part of the valuable lessons of the war. Much more important is the new attitude towards science that was engendered in the war, and the new integration that grew up between science and practical human activity. The war taught scientists not only to work but to think in a new way that can be summed up in the two words "infiltration" and "integration". As the war proceeded, the scientific workers found themselves less and less isolated and more and more part of a general directed effort. By coming closer to actual experience they learned the existence of problems at first hand and, also, the limitations which are inevitably imposed on types of solutions that can be used

in real circumstances by ordinary people. By ranging all the way from the front-line soldier, sailor or airman to the High Command, or from the worker at the bench to the Board of Directors, they appreciated, as they never had before, the relations of the functions of analysis and discovery to those of command and execution. At the same time that this process of infiltration was going on, the work of science was being integrated into an increasingly ordered effort to definable ends, and scientists themselves learned the virtues and the difficulties of effecting this organisation.

The full integration of science into war effort took place somewhat slowly and haltingly in this country. This was to a certain degree inevitable ; the mutual education of scientist and soldier could not take place all at once. But towards the end of the war, both in Britain and America, conditions had so improved that it is possible to see in them some approximation to a new and fruitful relation between scientific effort and the productive and operative forces of a modern state that goes far beyond the requirements and possibilities of war itself. The drive that led to this position was, of course, the direct polarisation of all effort to winning the war, and the increasingly intelligent understanding that this polarisation implied a full utilisation of resources under a system of well-thought-out priorities.

It is interesting to note that the scientists in the Axis countries did not enjoy these advantages ; for all the lavish assistance given to the scientific development of new weapons, the scientists were not trusted with the development and utilisation of their own ideas. There was no real infiltration of the military machine and little integration of scientific effort itself. All we are now finding out about science in Germany and Japan shows this waste of the capacities of scientists which is inherent in the essentially reactionary character of the Nazis and Japanese militarists, who both failed to understand and to trust their scientists.

We had in this country and, to a lesser degree in the United States, an absolute shortage of trained scientific men at all levels, from draughtsmen and laboratory assistants to theoretical research-workers. We therefore had to make the fullest and most effective use of those we had, and this led to a tacit acceptance of a principle which we should always have adopted, making man rather than money or equipment the measure of the effort expended. If anyone could be found with the capacity to analyse, initiate and direct scientific research and development, he was given full scope and money to do so, and, as we know from the achievements of the war, this was the secret of our great successes. It was most astonishing and invigorating to the academic scientist, cramped for years in a small laboratory and perpetually begging from this or that charitable fund for minute grants, to be allowed to spend thousands of pounds where he had previously spent ten pounds, and to engage assistance which effectively enabled him to give his time to the major problems and not to have to worry about gas leaks and typing. This scope resulted only in the rarest instances in extravagance or wasteful spending. The need to get the answers in the shortest possible time imposed an effective economy on expenditure. Planned experiments took the place of the old haphazard trials and demonstrations which had been usually more

expensive, and, because they almost invariably altered all the factors at once, provided no reliable answers. The freedom of scope for experimentation and assistance is a lesson which will not be lost on the scientists who experienced it. It will be of particular importance in the next few years when we are bound to suffer for lack of men to carry out the very much increased tasks which will have to be dealt with by science, and where it is more than ever important that we should make the fullest use of our few capable workers. The principle first enunciated by Professor Blackett, that allocation of money should be made to science according to what a competent scientist can usefully spend and not according to what he can just manage on, should be the basis for the finance of our post-war science.

Almost equally important as a lesson of the war is the value of the greater integration which was achieved in scientific work, partly through a more rational organisation and partly through the function of an effective positive information service. The organisation of science in the war service was naturally of somewhat gradual growth. Much of the earlier organisation tended to be bureaucratic and time-wasting, but with growing experience there grew up in the Service Ministries a close but flexible network of committees and panels manned by active scientists and in constant touch with service requirements. As time went on this network extended between the services and linked up with the scientific war effort of the Dominions and of the United States, but, most regrettably, never with that of the Soviet Union. This system provided what had been previously the function of the scientific societies, that is, careful discussion and interchange of scientific opinion, but it had also a much more positive function in laying out lines of attack and in determining priorities. In this way the scientific work itself could be carried out in a multiplicity of actual experimental stations, Government, Industrial, University, and yet not lose its coherence or general direction. Further, this direction was exercised by scientists themselves, at least in the latter part of the war, and was consequently sufficiently reasonable to be for the most part acceptable to the main body of scientific workers. Out of it emerged general concepts of organisation of science which will be of permanent value.

In the first place the distinction grew up between two kinds of studies of radically different characters—"object" studies and "subject" studies. In "object" studies the ultimate purpose was kept steadily in view, whether it was the defusing of bombs or the prevention of malaria. All the different approaches to the problem were correlated in the light of this object, and different sciences were called in to make specific contributions. "Object" committees had necessarily much closer association with the user than had occurred in peace time, but the user did not dictate the research or even put specific demands for equipment. As the war went on it became more and more possible to express what were known as Staff requirements in scientific terms, and to seek their achievement through research and development.

On the other hand it was often found that a large number of different objectives came to depend in some degree or other on a single field of science. This gave rise to the "subject" committees, which ranged all the way from the august Radio Board, which controlled the development of telecommunications and radar

for an enormous number of service users, to the humble "Mud" committee which considered the properties of mud in relation to service needs for civil engineering and transport. "Subject" committees were more homogeneous scientifically than "object" committees but, correspondingly, had a wider range of users. A simultaneous development of both systems of organisation are immediately applicable to peace-time science, and a great deal of the objections that have been raised to the organisation of science have rested on the fact that a distinction had never been made clearly between them. Many scientists had only learned to think of organisation as organisation with an object, usually an applied object, and were very alarmed lest the concentration on that object should prevent the natural, inherent development of particular scientific disciplines. But the parallel development of "subject" organisation should overcome this difficulty and, indeed, add very much both to the scope of scientific research and the number and novelty of problems presented by providing what has always been a notable gap in scientific organisation, a place where users and research workers can meet on the research workers' ground.

One other bugbear of the opponents of scientific organisation has been the danger of the dragooning of scientists which such organisation would necessarily imply. War experience has, on the whole, shown that this is not necessary. In fact, the most successful scientific efforts of the war were combined with the greatest freedom and scientific initiative on the part of the scientific workers. In one of the most important institutions in this country there was a regular weekly "soviet" of the research workers, who discussed the objectives as well as the details of the work they were doing. The pronouncements of the various atomic scientists groups in the United States shows to what extent a real democratic spirit was also present there. Where there was over-direction of scientists, it was usually because either the directors in question were not themselves scientists but were civil servants or military men, or because they were scientists of the old school of government science, and, in either case, it nearly always resulted in slow and ineffective working. In a few cases this became so bad that the system was changed, and the immediate beneficial results showed that democratic collaboration is at least as important in war-time as in peace-time science.

War science has often been objected to because of the secrecy it entailed, but, paradoxically enough, this is less apparent in war time because so many scientists are absorbed in war, and because the need for rapid communication of results makes is imperative for there to be a minimum of secrecy inside war science itself. This, of course, was not always realised : notably in the early years of the war and, very fortunately, all through the war in the Axis countries. By preventing the scientist from acquiring full knowledge both of the applications for which his work was required and of the parallel and contributory work of other scientists, initiative was hampered and even absurd results were obtained. It is always essential for a scientist to know why a question is being asked for him to give a sensible answer to it. The early stages of the development of atomic energy in the United States were seriously hampered by a security system which prevented one group of scientists from knowing what the others were doing. Towards the end of the war, however, these difficulties were overcome and a far more effective

system of scientific communication organised than most scientists had enjoyed before the war. Not only were there very efficient abstracting services and a free circulation of papers, but what was even more important, there was ample opportunity for scientists to go to discussions and to visit each other's laboratories and field stations all over the world. The advantages of air travel for the speeding up of scientific communication is something that scientists will not forget, and it is to be hoped that it will be equally available to them in time of peace.

However excellent the conduct of scientific research was in war time in detail, its value to the actual winning of the war would have been much smaller than it was if there had not been what Sir Henry Tizard has called a "general strategy" behind it. We scientists had much to learn from the traditional modes worked out through centuries of experience in the military field. The keen scientific worker who could attack and solve his problem was invaluable, but he was not necessarily capable, on account of his training, of seeing that the problem he was interested in solving was not the most important problem to be solved at that particular time, however technically interesting and absolutely valuable it might be. For that it was necessary to develop the type of scientific staff which, though it was never formally constituted, in fact existed in close connection with the General Staff in relation, not only to supply, but also to planning. This staff learned the value of evaluation of priorities and concentration of effort and, also, the even more difficult task of abandoning attempts even when they offered a fair chance of success, because the time for the use of them had passed. The strategy of war science did not exist in a vacuum. If nothing but scientific considerations existed, there would be hundreds of different ways in which attacks could be made on the infinite territories of ignorance that surround us. The choice must be determined, by and large, by short- and long-term human requirements. These requirements can easily be seen in war time as the achievement of victory. In peace, though less precisely formulable, they are those of the removal of want and fear.

The need for strategy in the application of science to war was the greater because of the disproportion between the demands of war for science and the supply of scientists. The total scientific effort of the country had to be considerably stepped up in an extremely short time. To a certain extent this was done by the more efficient employment of the scientists we had, which was made possible by the creation and operation of the Central Register. Once this Register came to be administered by scientists, it showed itself to be absolutely essential in the strategy of the scientific war effort in allocating appropriate workers to the different requirements. Naturally, the proportion of the requirements in the different sciences were not the same in war as they had been in peace. In the early stages of the war this led to important categories such as chemists and biologists being somewhat neglected, but later it was found that scientists were very much more interchangeable than their specialised training would have led us to expect, and that, particularly in the new branches of scientific war work, comprising Operational Research, scientists trained in all subjects could make equally significant contributions. The Central Register provides us with a new and, it is hoped, permanent addition to the armoury of science. By its intelligent

use it will be possible to predict sufficiently far in advance the needs in different scientific fields and, in relation to university programmes, to train the appropriate number of scientists. War experience has already taught us the need to keep our scientific man-power flexible by avoiding over-specialisation and inculcating general scientific method into teaching.

Even, however, when the most careful husbanding and wise distribution of trained scientists was effected, it was clear that we were far short of the scientific man-power requirements,* and here again the war has led to an innovation which we must try to preserve in peace. The system of science bursaries made it possible for the first time in our history to give to all those who had the ability to profit from a scientific education the opportunity to get one. The conditions, of course, were still not ideal. The demands of the fighting services and war industry naturally and rightly trenched on the supply of potential scientists, and consequently the standard for war bursaries was not as high as it might have been. But the system applied to higher education in general, and not only to scientific education, is one of the most valuable lessons of the war. Before the war, England gave higher education to only one and a half per cent of her citizens, and this one and a half per cent was not chosen on the grounds of ability. It is considered, on the basis of intelligence tests, that some eight per cent of the population would in fact be able to profit by university education, and it is most important that we should see that they get it and that the all-too-low standard of teaching should not be wasted by training those who have no better claim to education than the wealth of their parents. Even if this were done, the children of educated parents would have very considerable advantages, as they inevitably get additional education from a very early stage, which are not available to the children of most manual workers. It is interesting that the same problem has been recognised in the United States, where a far higher percentage—ten per cent—receive university education. In the Bush Report it is pointed out that this privilege is practically exclusively determined by parental wealth, and that there is an enormous waste of ability, which the Report recommends should be remedied by the promotion of national scholarships on a large scale.

The proper allocation of individual ability, to the tasks most suited to that kind of ability, has been very much advanced, from the experience of the Selection Boards which have operated in all the services. For the first time, scientific methods based on appreciation of psychological factors have been used on a really mass scale, and in so doing they have led to a much fuller appreciation of the relatively unknown factors involved. The selection of human beings is, of course, far more than sorting into categories ; it involves attitudes formed in family and social associations quite as much as any inherited trends, and it is therefore

* The United States had a very similar, though more intense, experience in mobilising scientific man-power. During the war the total national expenditure on research rose from about 350 million dollars, of which only some 70 million dollars were provided by the State, to not less than 800 million dollars, exclusive of atomic energy research expenditure. This vast increase did not, of course, mean that the number of scientists in the country had been doubled ; it only meant that they had been able, for the first time in history, to spend money freely on research. Nevertheless, the situation in America, where the expenditure on research was several times more than anyone imagines can be spent in this country, showed this still to be dangerously low. In the now well-known Bush Report, the demand for a much larger number of trained scientists is repeatedly made.

conditioned by social evaluations of different possible occupations. All these factors will have to be borne in mind in further application of selection methods in peace time, but the relatively rudimentary vocational guidance which existed before the war will undoubtedly gain enormously from war experience, and it is most important that it should become a permanent feature in the transmission of our culture to future generations. Another lesson of the war, closely allied to this, is the experience of training the comparatively uneducated personnel in the handling and maintenance of complicated scientific apparatus. The methods developed in training, particularly in the R.A.F., represented an enormous advance in the application of scientific methods to education. It is to be hoped that these lessons will not be lost, and will be extended to cover the whole range of peace-time scientific occupations, for it is certain that we shall need not only a much greater supply of university-trained scientists, but that a large proportion of the population will need to have practical mastery of the new appliances which are bound to be increasingly used in industry and in every-day life.

All the contributions of the war already mentioned are in some sense simply extensions of methods or ideas existing before the war. We have had organisation of science ; we have had selection and training for many years. The war has, however, introduced what is virtually an entirely new technique in the application of science. Operational Research, though having some affinities with certain commercial enquiries such as market research, arose independently in the war out of immediate needs, and developed so rapidly and so effectively that by the end of the war it was considered a necessary adjunct to every aspect of military affairs. At the beginning of the war new and essentially scientific devices, such as radar, were being prepared by the research departments of the services. To hand these over directly to service personnel in the field just would not do. Prejudice against new gadgets and unfamiliarity with their handling was combined often enough with practical unsuitability of the apparatus which had been designed by people of great scientific competence, but with little knowledge of the physical and human factors that occurred under service conditions. To remedy this state of affairs some of the scientists who had been actually concerned with the development were sent out to stations to supervise the use of the new apparatus. What might have been a mere servicing job grew into something much more important. For the first time immediate, personal liaison was effected between the scientist and the soldier, and the mutual education that grew out of it made the acceptance of further developments far easier. But it did more than this. The scientists in the field began to study the practical performance of their devices, and this study took on an increasingly quantitative and statistical aspect. From this study they began to deduce what were the real requirements of the devices as distinct from what they had been imagined to be by the military staffs, and this led at once to new directions in development. But the reciprocal process was also at work. Scientists began to see what the operational possibilities of new applications, which they either knew or could devise, were likely to be, and to influence operations in relation to these possibilities. As a very distinguished soldier said at the end of the war: "at the beginning tactics determined weapons, at the end of the war weapons determined tactics".

The first stage of operational research was concerned with physical devices, but once it had started it could not stop there. From the analysis of performance of particular weapons came naturally by extension the analysis of performance of a particular tactic, such as air-submarine chasing or the bombing of targets ; all the factors, and not only the instrumental factors, had to be brought in. Such factors involved many human elements : the morale or fatigue of the soldier or airman, their state of training, the reactions of the enemy; in fact, very soon, there emerged the beginning of military science in the modern sense, that is, a predominantly quantitative and statistical estimate of the operations of war. It is greatly to the credit of all ranks and branches of the services that this development, which could so easily have been resented as an intrusion, was on the contrary welcomed and encouraged. Before the war was half over Operational Research was already an accepted branch of operations, new and small to start with, but still taking its place beside the logistical and intelligence branches. It developed an *esprit de corps* and method of its own, and started vigorously recruiting among scientists of all kinds. It was found, for instance, that biologists and social scientists had a flair for operational research equal to, if not greater than, physicists or chemists.

Operational research led not only to greater understanding in detail of the operations of war, but to a much clearer integration of different types of operations. As the war went on, combined operations, whether by land and sea, land and air, or all three together, became the rule rather than the exception, and the bridge between the very diverse approaches of the different services was often effected through operational research. In this way several general principles emerged which have far wider application than merely to military operations. One of the most important was the intelligent use of variation analysis. The urgency of war makes it of cardinal importance that the improvements should be made first where the greatest return is to be expected for the least effort. By a study of a sufficient number of actual cases, where there has been a considerable range in variation in all the factors affecting the success of the operation, it was usually possible to select one that seemed the most promising, and then by deliberate variation to determine its effective return. At this stage it was often possible to introduce trials and purely physical calculations. These, however, could not be relied on implicitly because they inevitably omitted the operational factor; that is, the unreproducible character of the behaviour of men under the actual circumstances of battle. The neglect of this factor had, in fact, vitiated many of the earlier efforts at scientific analysis of warfare. What seems physically attainable, what may even be attained in trials, often proves entirely unattainable in battle. If it were not for the operational factor war itself would become an exact science. Indeed, it is not difficult in principle to analyse operations in such a way as to predict victory with certainty. If this could have been proved in practice, then war itself would cease to have any significance, for no one would engage in it on the certainty of being defeated. But, unfortunately, to reach such a result, some assumptions as to human characteristics had to be made, such as "that all Generals are infinitely intelligent". Operational research, dealing as it did with actual achievement, provided a quantitative corrective to

theorising. It also enabled the scientist to appreciate the human qualities of command and performance, and the conditions under which it is necessary to make and adhere to decisions where many of the vital factors cannot possibly be determined in time, or at all.

Operational research was concerned in the latter stage quite as much with the choice of problems to be attacked as in the attack on the problems themselves. It was clear that, however operational research groups were increased, they would never be able to deal with more than a small fraction of the soluble problems of warfare. Although it would always be necessary, in order to establish the value of the method, to deal with day-to-day problems or those specially brought up from the Staff, it was also necessary to have some guiding principles in planning the general use of operational research. The aim was to provide the most effective results in the shortest time. This determined priority of problems, on the basis that those which made the product of three roughly assessable factors the greatest should come first. The factors were the military importance of the solution, its tractability by any known scientific methods and the speed at which practical results might be expected. It would have been easy but wasteful to concentrate on the solution of many small but unimportant problems that could be handled quickly. It would have been equally wasteful to embark on vast schemes for reorganising the whole of warfare on a scientific basis. Further, small improvements were deliberately not looked for. At one stage in the war no improvement in the performance of any weapon of less than one hundred per cent within a year was worth considering. In peace the conditions of urgency will be less marked, but they will not be absent. All that will happen is that the time scale can be increased and the standard of desirable improvement lowered.

The knowledge accumulated in operational research and perhaps, even more, the experience of those who took part in it, may prove one of the most valuable contributions of the war to science. For operational research has clearly important applications to the post-war situation ; the same methods and, very often, the same men can be used—and some of them have been already—for the solution of peace-time problems in such fields, for instance, as building and the reconstruction of our export trades. We have found here a necessary link between laboratory science and industry.

One feature of operational research is particularly important in this respect. Operational research teams, as already mentioned, included social scientists ; in the new studies that will be needed as a basis for peace-time policy, even greater attention will have to be paid to the social aspects. For, complicated as are the human problems of war, there is in war an overriding general directive and an institutional discipline which is necessarily lacking in peace, where the extremely diverse and often conflicting desires of large numbers of individuals have to be taken into consideration. Here, the social scientists who have to find out what people want and need, and the physical scientists who hope to find the ways to satisfy those needs, will need to work together in the closest collaboration.

The original implications of operational research are already making themselves felt in peace-time economy. In principle it amounts to the statement that any human activity and any branch of that activity is a legitimate subject

for scientific study and, subsequently, for modification in the light of that study. Once this is accepted in practice, which implies the provision of research workers to carry out these studies, the way is open to a new level of man's control of his environment, one in which economic and social processes become scientific through and through. This is already happening in productive industry. We are witnessing what is really a new industrial revolution, in which statistical and scientific control and rational planning and design are taking the place that prime movers and simple mechanism did in the first industrial revolution. Industrial processes are now seen to represent cycles of performance, in which the needs of the consumer determine production, and are in turn modified by the results of that production, leading to a progressively greater degree of satisfaction at a steadily diminishing social cost. The stages of the cycle are also becoming clearer.

Beginning, necessarily, with the consumer, we have first of all the study of his needs; a study which is becoming more and more scientific, and where we are learning to distinguish between expressions of "wants", which indicate the attitude of the consumer in relation to his past experience and expectations and "needs", which represent the goods or services which will, in fact, give him satisfaction. The determination as a first approximation of these goods and services is a process in which we have learned much from the war. There the distinction between staff targets, where the general operational desiderata were stated, and staff requirements, when these were made precise in relation to actual possibilities of design and production, became very clear in the latter part of the war. Staff requirements can only be assessed if technical factors and production possibilities, as well as operational needs, are kept in view, and imply close integration between the sections responsible for determining requirements and the production sections.

To meet staff targets and staff requirements was the main task of research and development departments. These two aspects had tended before the war to be kept rather separate, and one of the lessons of the war has been to draw them together. Particularly in fields where rapid advances are an urgent necessity, as in radar or gunnery, it was necessary to expand these sections so that to a certain degree they acted as the major channel between requirements and production. Research and development are, however, essentially different modes of attack on practical problems. Research is concerned with the limitations of possibilities by material and known human factors; development has to take into account such factors as existing production facilities and economic considerations. In development, too, there enters the element of design where an apparently arbitrary selection is made between possibilities, but in which the specifically human and as yet unknown factors of skill and art comes in to supplement science.

After development comes production. At the beginning of the war there was, particularly in the service departments, far too great a separation between these phases of the productive process. It was soon realised that the actual designs submitted by the development departments, which were often very out of touch with modern productive processes, were most unsuitable for production as given

and had to be drastically modified—often seriously to their detriment—before they could be made in quantity. This difficulty was later overcome by the introduction of production engineers into development sections and by a much closer relation between the two departments. Quite apart from its integration with research and development, production methods themselves came to be influenced more and more by scientific considerations. One of the great contributions of the war to industry was the introduction of statistical quality control and, with it, the possibility of far better quantitative planning of production processes. In other words, it was recognised that the production process itself could be scientifically studied and scientifically controlled.

The final phase of this productive cycle, involving the examination of the performance of the products and the degree to which they satisfy the original requirements, has already been discussed as a starting point of operational research. It merges completely with the first phase of assessment of needs, and requires the same survey techniques. In peace time, once it can be divorced from the elements of commercial competition, and so be free from the suspicion of deliberate bias, the subsequent servicing and customer studies will be a most valuable guide to the continuous improvement of production. These phases of the scientific control of the production cycle are clearly interdependent, but more is required than the necessary liaison between people carrying them out. There must be some mechanism, such as was provided in the war by the scientific directorates and their advisory committees, and which has begun to appear in peace as "industrial working parties", to determine the major lines of policy, and to allocate the scientific effort according to where it is most likely to be useful.

In the development of the atom bomb we have had an example on the largest scale and the maximum intensity of scientific enterprise on the new model. There, pure scientists, engineers, adminstrators and soldiers all worked together in a coordinated way according to a definite plan, flexible enough to allow for the necessarily unforseen contingencies of scientific discovery. In some ways this organisation was as important an invention as the liberation of atomic energy itself. The lesson that it teaches is that it is possible, given a definite end and a social will to achieve it, to attack any problem and to reach practical success in a small fraction of the time in which it might have been reached by the uncoordinated methods that ruled before the war. Actual speed of advance will be limited in the first place by the amount of scientific and technical man-power, together with their accompanying material resources, that are put on the job. Only when this exceeds a certain value is progress likely to be held down to the limits fixed by the obstinate and unalterable properties of material systems, or the almost equally obstinate difficulty in the solution of intractable intellectual problems.

The choice, whether to push matters to these limits, is a social choice : it depends on an overall assessment of what tasks the community most needs carrying out, but this in turn, within limits, is a scientifically determinate question. Choices within the general direction of the scientific war effort had, indeed, constantly to be made during the war. In the early phases they were frankly arbitrary choices by eminent politicians, guided more or less by self-chosen

or official advisers ; towards the end, a more scientific organisation tended at least to modify this arbitraryness. But certain arbitraryness must and will remain ; there often comes a point when it has to be recognised that we lack the knowledge to make an important decision and that someone will have to make the decision before we can possibly hope to get that knowledge. The function of a scientist as a scientist does not extend into this field. Responsibility for making decisions which are basically indeterminate and must take unassessable factors of human attitudes into account, is the proper field of executive authority. The scientist is, however, responsible for seeing that executive authority is fully aware of the assessable factors and of the consequences of different decisions, in so far as they can be known, and it his duty, not only to make this information available at request, but to force it on those in control by all the means in his power. There was a saying in the services that the scientist must be on " tap " but not on " top "! Now this, although apparently plausible, contains a very dangerous simplification. The scientist does not claim or want to be on top, but it is not nearly enough to have him merely on tap. In most cases the executive authority will not be able to see for himself when the scientist should be called in or what questions he should be asked.*

The war has brought a terrific increase in our positive means of control of the material environment, culminating in the discovery of the utilisation of atomic energy. It has brought a hardly less important advance in the possibilities of human organisation. In both directions, these advances mark, at first sight, only an acceleration of processes already well in evidence before the war. It is now becoming more and more obvious that the total magnitude of these changes is producing what is really a new stage in human development, a stage where the conscious integration of human effort is taking the place of development dependent on the unplanned interactions of millions of human wills, often themselves unconsciously motivated. The year 1945 may well be taken to mark an era of ultimate importance greater than that of any transformation since the beginning of human society. Once human society had appeared on this globe, bringing with it the possibility of cumulative tradition and progressive alteration of society itself, the old processes of organic evolution that produced the myriads of species, painfully adapted through natural selection for their mode of life, were completely outpaced in world development. Evolution took millions of years to change things which human beings could do in thousands. Similarly, the new, scientific, conscious, social organisation, once it is fully established, will make the older forms of complex historical development equally obsolete and, at the same time, enlarge to bounds, the limits of which we cannot conceive, the possibilities of material and social advancement of mankind. Naturally, many who have just glimpsed at this possibility, but who have not

* On one occasion in the war, a scientist discovered that an operation was going to be undertaken which would, if carried out as planned, result in certain loss owing to physical conditions. He managed with considerable difficulty to get this recognized and acted on, and only later discovered that a report to the same effect presented by his predecessor had been lying in the files of the Command for several months. According to the Services code the latter scientist had done his duty. If the matter had been in the field of science he would have established priority of publication, but in the field of action the battle might have been lost, because of his nice sense of the limitation of the scientist's function.

understood it—or only understood it in terms of the past—shrink from this opportunity and the immense responsibility which it places on man for the conscious direction of his own future. They are in the position of the wild men of the woods of the previous transformation, who preferred to fend for themselves as had their animal forebears, rather than mix in the dangers of the disturbing affairs of human society.

The problem before us now is how to utilise the lessons of the war for peace—not only the lessons that have been outlined above, but the infinitely greater number of lessons which I have had neither the knowledge or the space to deal with, but which are embedded in the experience of those scientists who have taken an active part in the war effort. Most immediately, there are, of course, the direct technical applications of the scientific advances of the war—in physics, chemistry and medicine; actual factual results of war gained knowledge, and the use of apparatus developed in the war can be turned to immediate advantage. I will cite only one example from my own experience in addition to that already mentioned on page 419. The operations of the intelligence and the engineering sections in the great variety of theatres in which the war has been fought, has led not only to an enormous increase in our geographical knowledge but, even more, to a revolution in geographical methods. The principle underlying these is the use of air photography, the methods of which have been improved out of all recognition during the war, and which can be applied immediately to hundreds of problems of academic and economic geography in the peace. Aerial photography makes it possible, for the first time, to hope for a reasonable topographically accurate map of the whole surface of the globe within a few years. Its value for geological and ecological studies is also manifest, and its oceanography value is hardly less. Combined with new methods for measuring atmospheric and sea surface conditions, including waves, we may be able to make enormous advances in our understanding of the interrelations of sea and land. During the war, for practical purposes, investigations on such processes were carried out by means of integrated, full-scale and model experiments. We could now take up again the study of erosion and transport of materials where the Victorian amateurs left it and, with the use of explosives, bulldozers and other modern heavy equipment, make the beginning of the study of experimental geology.

What could be done for the science of the earth could be duplicated over nearly every other field of science. The individual scientific research worker has learned through his war experience that studies which he had hitherto considered impossible were merely impracticable because of lack of resources. The lesson will be wasted unless he is in peace enabled to continue to use and even expand the use of similar resources. Further, the war has given the possibility of quantitative justification of such research. Many scientists may have felt before the war that if they were allowed to do certain research the community would benefit, but they could never prove it and not enough were sure of it to carry conviction. Now in practically every field of science he can point to assured results from investigations properly planned, adequately operated and fully supported in equipment and assistance. The productivity of science

is now a measurable factor. The difficulties we foresee in the future are not so much those of lack of support as lack of men to carry out the work.

We know that for any particular project, a resolute attack at an adequate scale of effort will bring a fairly certain return ; naturally, as we are dealing with the unknown, we cannot always ensure that the object achieved is the object that was at first envisaged, but this does not matter if the object achieved is even statistically—that is, over the average of all ventures—equally valuable. A research into the cure of a particular disease may fail, but there is a very high probability that it will provide a cure for some other one or provide an addition to pure or applied science in another field. Knowing this probability of success, however, does not mean opening an indiscriminate number of new crusades into the unknown ; by pursuing this or that object with whatever fair hope of success we may be diverting resources from others more important or more promising. The strategy of scientific advance in peace is no less important than it was in war. By a survey of the whole field of the fundamental and applied sciences, and by the analysis of that survey, it should be possible to indicate at least the major directions into which scientific effort might be directed with some idea of their quantitative importance.

Now this raises the old issue of planning and the freedom of science. It is to be hoped that the controversy is dying out as experience begins to show its unreality. The war showed that science can be planned successfully, but it also showed that success was only achieved if the scientists concerned were working in that full freedom of action and at a job about which they were themselves enthusiastic. The skill of the organiser of science in war was very largely in fitting the jobs to the people, and those conditions are not substantially changed in peace. We still have far too few scientists of ability to dream of wasting one of them in directing him to work on something in which he is not vitally interested. Planning does not mean arbitrary dictation from above : the ordering of people to solve this or that specified problem. It does mean that the scientists themselves must get together to find out what their problems are and to set about in a flexible but vigorous way to solve them.

Now, of course, this cannot be done without some organisation, and the type of organisation is becoming much clearer in the light of war experience. Many practical people, and some of the more confirmed anti-planners, imagine that the planners wish to attach each particular science to the service of a particular industry—the physicists to the electrical industry, the chemists to the chemical industry, the biologists to agriculture or medicine—but those who have studied the interrelations of science in practice know that it is not as simple as this. Every cumulative discipline of science, with its own tradition, its own grammar and vocabulary, cannot be so isolated and attached. Physics, for instance, not only penetrates all other branches of fundamental science but affects every industry and agriculture and medicine as well. At the other end of the science spectrum, psychology has an equally wide sphere of utility. And the converse is also true. Not a single industry, hardly a branch of any industry, but calls on the service of every discipline of fundamental science—naturally not all to the same extent and some more directly than others. The detailed problem of the organisation

of science is the problem of linking these multiple sources to these multiple ends, so as to provide the greatest degree of freedom and flexibility. As a provisional solution, opinion is coming round to using an intermediate station as a kind of exchange control between fundamental and applied science : what are coming to be called "background" scientific institutions, such services as have been provided, for instance, by many departments of the National Physical Laboratory. The "background" institution covers a group of sciences on one hand and industries on the other. For instance, a background institution on silicates would draw mostly on physics, chemistry and geology and serve the metal, ceramic and building industries as well as civil engineering and agriculture.

In such an organisation, which itself deserves to be scientifically studied and not allowed to grow up haphazard, a proper balance would have to be maintained between fundamental, background and applied science. All those who realise that there is no simple relation between science and its application will see the necessity for maintaining vigorous, independent, fundamental science on an absolute and relative scale far greater than ever before. One lesson that the war has taught us is that from every fundamental science may come—as in radar and the atomic bomb—practical, useable applications in a time measurable in months; on the other hand the stimulus of practical needs, equally rapidly transmitted, may provoke the discovery of new problems in fundamental science. To maintain this two-way flow rapidly and effectively is one of the main practical problems of the new organisation of science. The general objective is that of the positive development of scientific ideas in the light of scientifically assessed human requirements.

To achieve it we need, in the first place, a far better system of communications than we have yet enjoyed in peace time. The system of scientific publication which was in existence before the war was already admittedly completely overloaded and was beginning to break down. Since then it has been largely wrecked as a result of the war. The German scientific publications that have disappeared and the various indexing and summarising functions undertaken in Germany, will not be replaced, and meanwhile shortages of paper and labour have reduced the possibilities of exchange of information to a position worse than anything that has existed since the Middle Ages. Even if full recovery is made from these conditions, the major implication of a growing science in an increasingly scientific society will have to be faced. The quantity of ascertainable facts, and the rate at which facts are ascertained, will continually increase, while the quantity assimilable by any scientific worker is absolutely limited and is already near its limit. The way to solve this problem is clearly to devote some time, money and man-power to a properly recognised system of communications, which could be one very largely based on the experience of wartime science. Such a system, however, will not arise of itself. The danger is that scientists will be so preoccupied with their own problems that they will accept inadequate communications, and thus hamper their own work in a way which is no less serious because it may not be recognised. The need for the restoration and permanent improvement of scientific communications is one of the legacies that the war has left to science, and one which can be dealt with only by scientists themselves.

The general character of post-war science, particularly that under Government control, will differ from pre-war science in one very important respect. The war has shown us that controlled scientific development is an essential part of the full utilisation of scientific discovery ; this implies that the material equipment of new governmental establishments should be very much more on factory or semi-factory lines, or, in fact, like those of Service establishments or similar establishments, in peace. Emphasis on development will also serve to bring together the scientist and the engineer with mutually valuable results, and will tend to promote a much freer interchange of career, taking the place of the rather rigid demarcation of before the war.

The most balanced and flexible plan for scientific research, however, will not be enough. It will need to be integrated with a positive drive : a technical, biological and social advance carried out with all the resources of the community. That such a task can be achieved has been shown by the experience of the war ; but the war has shown, also, that it is not only possible but absolutely necessary for survival as an advanced community. A national economy, integrated through science and continually advancing by means of scientific research and development, is the basic need of the new era which we are now entering. It implies the expenditure of a much larger proportion of social effort and social resources on science than ever envisaged before.

Those who had considered the advantages that science could bring to society had realised well before the war that the expenditure on science by society was far too small. At that time the total expenditure in this country was something of the order of one-tenth of one per cent. of the national income ; they could see, and they tried to point out, how the increase of this proportion would bring far more rapid prosperity. In the post-war situation, however, with the leeway of destruction and disorganisation to make up, and the far weaker and even perilous situation of this country, what was a desirable end has become an absolute necessity, and the proportion to be aimed at must be a much higher one. A committee of the Lord President of the Council is now engaged in assessing the country's requirements of scientific man-power. With all the official data at their command they can obviously make more exact short-term assessments of the needs and possibilities than can ever be made from outside, but it is fairly safe to assume that whatever they recommend it will imply that the fastest attainable rate of increase of scientific man-power will still fall short of our real needs. On the long-term view we must look forward to a fairly rapid transformation in which scientific functions—not necessarily scientific research and development only, but scientific production and scientific administration—will absorb a progressively larger and larger proportion of the population. From one-tenth of one per cent. we may advance to involving one, two and possibly, ultimately, but in the far distant future, as much as twenty per cent of the population in such activities. This is a logical consequence of the increasing rôle of human intelligence and consciousness in the management of our society. Long before such a stage is reached, however, the distinction between scientific and non-scientific activity will probably have largely disappeared. Already we require for the proper functioning of our society a certain degree of knowledge

of the facts of science and even more of its method on the part of every citizen. The people cannot direct the Government to make wise decisions, nor can they carry out the decisions reached, unless they have much fuller understanding than at present of what they are doing.

These considerations necessarily raise the question that is already becoming acute : the position and responsibility of the scientist in society. The recent controversies on the atomic bomb and the degree to which a question—a fundamental one of natural science—has become the central point of dispute in international politics, has forced governments and politicians to pay attention to scientists and, at the same time, scientists to concern themselves as never before with political questions. Rough language has been used against the alleged threat of scientists usurping the functions of the State ; scientists on their side have reasserted their responsibility for the maintenance of truth and their demands for the removal of restrictions on its spread. On reflection the spheres of responsibility of the scientist and the politician are seen to be distinct, even though they have come to impinge on each other to a greater degree than ever before, and will certainly interact to an even greater degree in the future. It is now a fact that no step can be taken on any serious social or political topic, whether it be trade or industry, administration or war, without involving intricate technical and scientific questions and, correspondingly, hardly any discovery of importance can be made in science without important political implications. Now, clearly, if the statesman does not understand far more fully than he has ever had to before what is the nature of these discoveries, he will not be able to act reasonably in dealing with them, and conversely the scientist, if he knows as little about politics, will not be able to point out to the statesmen what the consequences of these discoveries are likely to be in the world of affairs. That does not mean that the scientists claim to be the government or to have any special statutory position. The responsibility for final decisions necessarily taken on incomplete knowledge is the business of those selected by the people to act on their behalf. The scientist's job is only to see that the information is as complete as it can possibly be at the time, and that research is being adequately carried out to ensure the acquisition of more accurate information.

This naturally implies a more responsible rôle for the scientist than he was conceived to fill before the war. The conception then was that the scientist in the government service was a consultant to be asked definite questions referring to technical matters but not to be concerned with formulating these questions or the policy that resulted from the answers he gave. Such a limited function was already proved to be quite inadequate in the war and will be even more so in peace. In the first place, it is self-defeating to limit the subject of enquiry to technical matters : all relevant factors—social, economic and administrative, as much as technical—need to be known before adequate answers can be given. In the second place, an answer given without any relation to the intention for which it was asked is already hopelessly limited and often biased by ignorance. In war it was often more important to know why a question was being asked than what the question was. To work effectively in the governmental or industrial

machine the trusted scientist must have the fullest access to all sources of information and must be an equal member of all important bodies at which decisions are made. This, not so much because of the importance of his vote, but more because the need to be "in the picture" is an essential prerequisite to the formulation and solution of practical problems. This in itself, however, will not be enough. It is also essential that the scientist should have the power and the means to initiate fact-finding research on all matters on which decisions have to be taken and to ensure the results of this research are published.

These are perfectly reasonable requirements that have the precedent of the war behind them. They would, of course be much more rapidly granted if, on the other side, Parliament, Government and the Civil Service had more appreciation of the nature of scientific method and the outlines of scientific fact, but we cannot afford to wait for the necessary change in our educational system for this to happen. Actual prejudice apart, it is by no means impossible through the experience of collaboration between scientists and administrators to build up a mutual understanding. At the same time organisational changes will be necessary. It is to be hoped that these will be recommended by the Committee on scientific man-power. To make administration adequately scientific requires both a stronger central scientific body to deal with general questions and a more widespread system of scientific advisers in all the departments of state.

The changes proposed, though radical, we now know are both possible and necessary, for on them depend the successful transition from a pre-scientific and planless economy to a fully planned one with the minimum of loss of efficiency and liberty. It is perhaps not fully realised how much science itself can do in removing the arbitrary and despotic elements which many persons of genuine liberal feeling imagine to be inherent in all planning. Acts are arbitrary in proportion to the lack of knowledge of the situation with which they are made to deal. The more scientific the analysis of the situation the more possible it is to find a solution which will really answer the need and at the same time be the most acceptable one as demonstrably reasonable. Science, by accepting corporate opinion and reason as its criterion, is itself a democracy : one always open to conviction but not accepting any dictum until it has been convinced. In so far as science infuses government, it enhances all the democratic elements in it.

All these considerations, which have arisen from the first stage of appreciation of the enormous scientific effort that went to the war, will, no doubt, be amplified and reinforced when these lessons are more fully digested ; but we have enough already to be able to show how the scientist can turn to advantage in the welfare of man the energies and methods which for the past years have gone to his destruction. The war has transformed the world, has left much destruction and many anxieties, but it has also given us a demonstration of the possibility of the rational removal of these evils and their replacement by a better state than we have ever before been able to imagine. That is the major lesson of the war for scientists.